APPLIED BIOINFORMATICS FOR EXPLORING DIVERSITY PATTERNS IN META-OMIC DATA

A Dissertation
Presented to
The Academic Faculty

by

Piyush Ranjan

In Partial Fulfillment
of the Requirements for the Degree
Master of Science in Biology in the
School of Biological Sciences

Georgia Institute of Technology
May 2018

APPLIED BIOINFORMATICS FOR EXPLORING DIVERSITY PATTERNS IN META-OMIC DATA

Approved by:

Dr. Frank J. Stewart, Advisor
School of Biological Sciences
Georgia Institute of Technology

Dr. Jung H. Choi
School of Biological Sciences
Georgia Institute of Technology

Dr. Jennifer Glass
School of Earth and Atmospheric Sciences
Georgia Institute of Technology

Date Approved: April 20, 2018

To my beloved Grandparents

ACKNOWLEDGEMENTS

I would like to express my deepest gratitude to my thesis advisor, Dr. Frank Stewart, for his immense support and guidance throughout my time working with him and his lab. I will always appreciate his time whenever I had concerns and questions about my research or writing. Besides him, I would also like to thank Neha Sarode, Zoe Pratte and other previous and current members of the Stewart lab for their helpful discussions and suggestions about my research.

I am grateful to our collaborators, Dr. Gang Bao and Dr. Jennifer Glass, with whom I have had the pleasure to work on meaningful scientific research. I would also like to acknowledge my dissertation committee members, Dr. Jung Choi and Dr. Jennifer Glass, for their encouragement, insightful comments and hard questions.

I am thankful to my friends, Dr. Vinay Mittal, Siddharth Choudhary and Dr. Lavanya Rishishwar, for their great company, unfailing support and continuous encouragement throughout this time.

Nobody has been more important to me in the pursuit of this research than the members of my family. I would like to express my profound gratitude to my parents for their love, guidance and trust throughout my life. Most importantly, I must acknowledge my wife and best friend, Shipra, for her love, support and assistance, without which I could not have accomplished what I have today.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS AND ABBREVIATIONS

AMZ	Anoxic Marine Zone
Cas	CRISPR Associated Protein
CCR5	C-C chemokine receptor type 5
CRISPR	Clustered Regularly Interspaced Short Palindromic Repeats
CSCG	Core Single Copy Genes
gRNA	(small)guide-Ribonucleic Acid
HBB	Hemoglobin subunit beta
HBD	Hemoglobin subunit delta
HRD	Homology Directed Repair
MAG	Metagenome Assembled Genome
NHEJ	Non-Homologous End Joining
OMZ	Oxygen Minimum Zone
OTU	Operational Taxonomic Unit
PAM	Protospacer Adjacent Motif
RGN	RNA-guided Site-specific Nucleases
SAG	Single-cell Amplified Genome
TALEN	Transcription Activator-like Effector Nuclease
ZFN	Zinc Finger Nucleases

SUMMARY

This thesis explores the utility of applied bioinformatic approaches to better understand sequence space and phylogenetic diversity in meta-omic clinical and environmental datasets. In three chapters, the thesis describes how applied bioinformatic techniques can be used to 1) identify and quantify sequence variation in the form of insertions and deletions generated as an effect of off-target activity by CRISPR-Cas9 nuclease using high throughput targeted gene amplicon sequencing; 2) identify and quantify the abundance of elements of the bacterial defense systems, CRISPRs, to explore viral-microbe interaction dynamics in natural microbial communities living in marine oxygen minimum zones using high-throughput metagenome sequencing; and 3) investigate phylogenetic variation in an underexplored phylum of bacteria, Atribacteria, that are found as dominant members of microbial communities in methane hydrate-bearing marine sediments again using high-throughput metagenome sequencing.

CHAPTER 1. IDENTIFICATION AND QUANTIFICATION OF INDEL VARIATIONS GENERATED BY CRISPR-CAS9 NUCLEASE OFF-TARGET ACTIVITY

1.1 Introduction

The aim of this research was to develop a bioinformatic method for automatic, high-throughput quantification of indel mutations generated because of CRISPR/Cas activity through the use of next-generation sequencing. This research was a part of a greater effort towards understanding off-target activity of CRISPR/Cas mediated gene editing in human cell lines. To achieve this goal, our collaborators performed experimental perturbations varying the components required in the CRISPR-Cas9 activity. The experiments were sequenced on a next-generation sequencing platform and were analyzed using the proposed bioinformatic method. This research became a part of the publication Lin, Y., Cradick, T. J., Brown, M. T., Deshmukh, H., Ranjan, P., Sarode, N., ... & Bao, G. (2014). CRISPR/Cas9 systems have off-target activity with insertions or deletions between target DNA and guide RNA sequences. Nucleic acids research, 42(11), 7473-7485.

1.1.1 CRISPR-Cas system as microbial adaptive immunity

Viruses are hugely abundant in the environment (Suttle, 2005) and key factors in shaping the ecology and evolution of all branches of life, bacteria, archaea and eukaryotes, owing to their ability to act as a predator as well as a vector for genetic exchange (Chibani-Chennoufi et al., 2004). To fend off these predators, every host has been observed to demonstrate some form of defense that can generally be grouped as innate and acquired

immune systems. Innate systems, for example the restriction-modification in microorganisms, are non-specific and often target generic features of the infection. Adaptive or acquired systems, for example B and T cells in humans, have the ability to learn and protect the host from subsequent infection by recognizing specific features of the pathogen. While adaptive immune systems are fast, efficient and capable of mounting a defense in the same generation, they are evolutionarily complex to develop. Innate systems, however, are simple and act as a first line of defense but are shaped by constant evolution of the host making them less effective than adaptive systems in a generation.

Bacteria and archaea, dominate almost all natural and artificial habitats, including inhospitable environments, despite predatory viruses always looking for opportunities to decimate their populations. To survive, they have developed a variety of natural defense mechanisms to fend off invaders (Labrie et al., 2010). Many defense systems target diverse steps of the viral life cycle, notably blocking adsorption, preventing DNA injection, restricting the incoming DNA, and abortive infection systems. These systems are generic in nature, and abundant owing to their simplicity and work as innate immunity systems. The idea that an adaptive immunity system, because of its complexity, could be a part of a prokaryotic genome was uncommon until 2007, when Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR) loci in *S. thermophilus* were shown to acquire novel spacers derived from invasive phage DNA (Barrangou et al., 2007). Since then, CRISPR-Cas modules have been described in most archaea and many bacteria (Deveau et al., 2010; Marraffini and Sontheimer, 2010; Koonin and Makarova, 2013; Makarova et al., 2013; Barrangou and Marraffini, 2014). The CRISPR locus was first observed in *Escherichia coli* in 1987 (Ishino et al., 1987) and then in 2002 (Jansen et al., 2002) but its biological role

was not well understood until the spacer sequences were reported to be homologous to foreign genetic elements including viruses and plasmids in 2005 (Bolotin et al., 2005; Mojica et al., 2005; Pourcel et al., 2005). However, the full realization of the potential of CRISPR-Cas system as a RNA-mediated, sequence specific defense system was observed after subsequent studies between 2006 and 2008 (Makarova et al., 2006; Barrangou et al., 2007; Brouns et al., 2008).

CRISPR, which stands for Clustered Regularly Interspaced Short Palindromic Repeats, and CRISPR-associated (Cas) proteins provide sequence-specific protection by cleaving foreign DNA or, in some cases, RNA (Horvath and Barrangou, 2010; Garneau et al., 2010; Hale et al., 2009). A CRISPR locus consists of a CRISPR array, comprising short, partially palindromic DNA repeats that are present at regular intervals and form loci that alternate repeated elements (CRISPR repeats) and variable sequences (CRISPR spacers), which are flanked by diverse *cas* genes. These *cas* genes encode proteins essential to all the steps in which CRISPR-Cas system mounts the immune response (Barrangou et al., 2007; Brouns et al., 2008). CRISPR spacers are often derived from nucleic acids of infecting virus and plasmids and are used towards recognition of matching DNA and RNA from similar sources to destroy them. Since the CRISPR array in the genome is modified during the spacer acquisition, this protection can be inherited. This array modification usually happens at one side of the CRISPR cassette, which also makes it a chronological record of the infection experienced by the host and its ancestors.

CRISPR-Cas immune system functions in three stages: adaptation, expression and interference (Figure 1). These stages have been discussed extensively in several research articles since 2008 (Barrangou, 2013; Barrangou and Horvath, 2012; Fineran and

Charpentier, 2012; Marraffini, 2013; Reeks et al., 2013; Sorek et al., 2013; Westra et al., 2012; Wiedenheft et al., 2012; Garneau et al., 2010; Magadán et al., 2012). During the adaptation stage, fragments of exogenous DNA, better known as protospacers (Deveau et al., 2008), are incorporated in to the CRISPR array as new spacers from the invading viruses and plasmids. These protospacers are immediately flanked on one side by a highly conserved CRISPR motif known as protospacer adjacent motif (PAM), which is usually 2-5 nucleotides long (Mojica et al., 2009). PAMs have been known for their involvement in both spacer acquisition and target cleavage (Paez-Espino et al., 2013; Gasiunas et al., 2012; Jiang et al., 2013; Jinek et al., 2012; Sapranauskas et al., 2011; Sashital et al., 2012). Mutations in the PAM motifs have been implicated in developing CRISPR-resistance by viruses (Semenova et al., 2011; Sun et al., 2013). Spacers inserted into the CRISPR array demonstrate the memory acquired from each infection and later act as sequence templates for a targeted defense against subsequent invasions by the corresponding virus or plasmid.

Cas1 and Cas2, which are metal-dependent endonucleases (Beloglazova et al., 2008; Nam et al., 2012; Samai et al., 2010), are present in most known CRISPR-Cas systems and form a complex representing an adaptation module. This adaptation module enables insertion of spacers in the CRISPR array (Nunez et al., 2014; Yosef et al., 2012). The adaptation stage is followed by the expression stage, where the whole CRISPR array is transcribed in to a precursor transcript (pre-crRNA). The pre-crRNA is bound to either Cas9, which is a single multi-domain protein, or to a multi-subunit complex forming the crRNA-effector complex. Individual mature CRISPR RNAs (crRNAs) are processed via an endonuclease activity on crRNA-effector complex. Depending on the type of CRISPR-Cas system, this fragmentation can be achieved through an endonuclease subunit of the

multi-subunit effector complex (Wang et al., 2011) or an alternate mechanism involving a subunit of Cas9, bacterial RNase III and an additional RNA species, the tracrRNA (transactivating CRISPR RNA) (Deltcheva et al., 2011). The mature crRNA thus produced is used in the interference stage towards inhibition of foreign genetic material. During the interference stage, the mature crRNA remains bound to the Cas9 or to the multi-subunit crRNA-effector complex and acts as a guide binding to homologous DNA or RNA sequence from a viral or plasmid origin. This binding triggers the nuclease activity of Cas proteins which cleave off the cognate sequence (Samai et al., 2015; Hale et al., 2012; van Duijn et al., 2012; Zhang et al., 2012; Wiedenheft et al., 2011; Makarova et al., 2011; van der Oost et al., 2009). A potential common ancestry has been suggested between the CRISPR-Cas system components and the components of other defense mechanisms like restriction modification and toxin-anti-toxin systems as well as mobile genetic elements like transposases (Makarova et al., 2013).

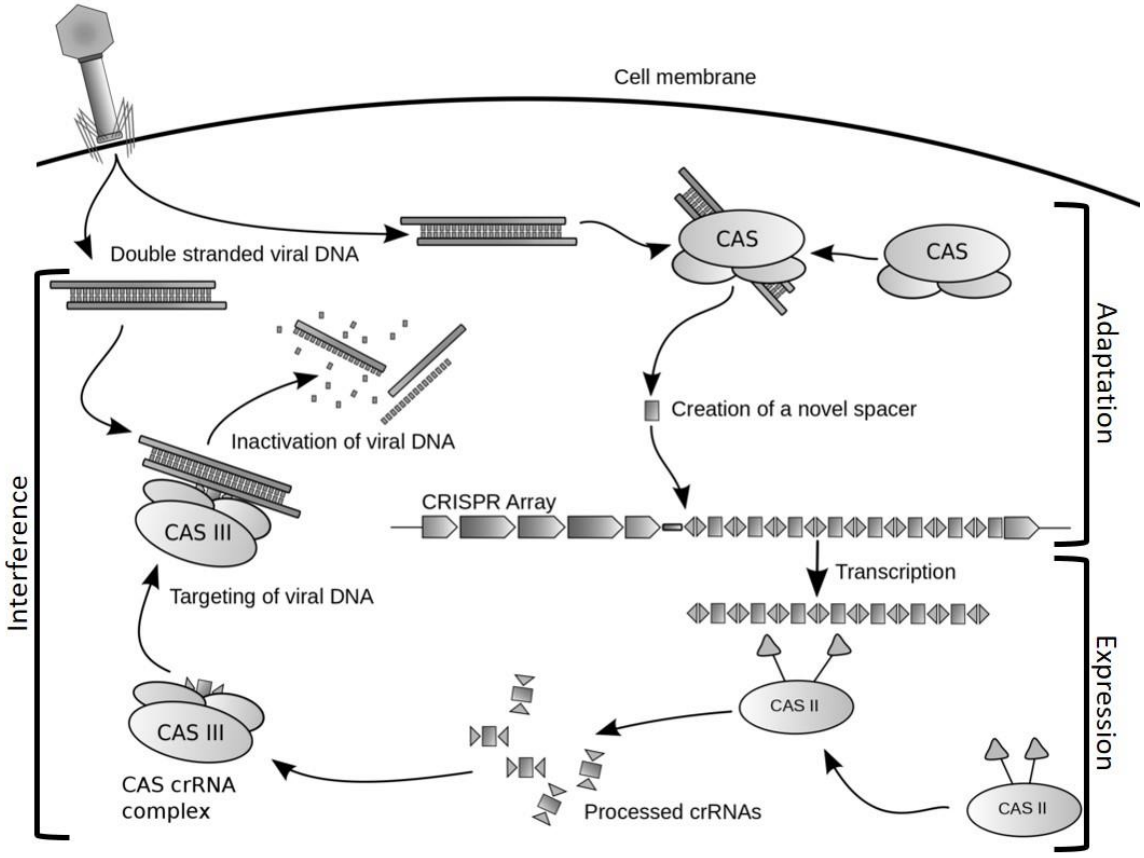


Figure 1 – Schematic action of a bacterial/archaeal CRISPR-Cas mediated adaptive immunity.

CRISPR loci contains an array of spacers (rectangles) surrounded by repeat elements (triangles), flanked by a leader sequence and cas genes. CRISPR-Cas mediated immunity works in three stages 1) Adaptation: acquisition and insertion of new spacers in the CRISPR array, 2) Expression: transcription of CRISPR array and processing of crRNA and 3) Interference: detection and degradation of foreign genetic material by crRNA-Cas complex. Modified from Atmos, 2009.

1.1.2 CRISPR-Cas systems as gene editing tools

Technologies that enable DNA manipulation *in vitro* and *in vivo* have facilitated several major advances in the field of Molecular Biology. Purification of bacterial restriction endonucleases that cleave DNA at specific genetic loci and later the development of recombinant DNA techniques that combined the use of restriction

endonucleases with DNA ligases and polymerases marked the starting of gene editing. Instead of relying on selective breeding, gene editing opened a faster and efficient channel, for example, to enhancement of plants and animals that have a better yield and are more resistant to diseases. Similar enhancement has been seen in scientific research where gene editing has enabled creation of user-defined cells, model animals and gene-modified stem cells with novel characteristics.

Before CRISPR-Cas systems were established as a method for precision gene editing, *in vivo* genetic engineering was dependent on technologies that involved zinc finger nucleases (ZFNs) (Carroll et al., 2011) and transcription activator-like effector nucleases (TALENs) (Boch et al., 2009). These techniques worked by using a nuclease enzyme that introduces a double-strand break at its target site along with a protein engineered to bind to a specific DNA sequence. Although, the DNA cleavage is accurate, these techniques require proteins specific for their target sites. RNA-guided site-specific nucleases (RGN), on the other hand, are a combination of a nuclease along with a guide RNA for specific binding (Marraffini, 2015; Lander, 2016). CRISPR-Cas systems, as a RGN, may provide an exciting alternative to ZFNs and TALENs as the nuclease remains the same for different gene targets while the short sequence of guide RNAs can be readily modified to redirect the site-specific cleavage (Cong et al., 2013). This opens possibilities for DNA editing at several targets simultaneously and at any locus in a genome.

CRISPR-Cas systems have the ability to introduce double stranded breaks in DNA (Figure 2). These breaks are an important feature of gene editing tools that make use of endogenous DNA repair pathways in eukaryotic cells to make genetic alterations through either Non-Homologous End Joining (NHEJ) (Moore and Haber, 1996) or Homology

Directed Repair (HDR) (Pardo et al., 2009; Bolderson et al., 2009). During NHEJ, blunt-end cuts are repaired by direct joining of the two fragments, often deleting a few bases and rendering the gene ineffective by either damaging the gene product or causing a frameshift mutation. During HDR, the damaged DNA is repaired with a homologous DNA fragment usually available as another allele. This property of HDR can be leveraged for any type of gene modification through insertion of a homologous DNA fragment along with the delivery of the cleavage method, which in this case could be the CRISPR-Cas system (Cong et al., 2013; Mali et al., 2013b).

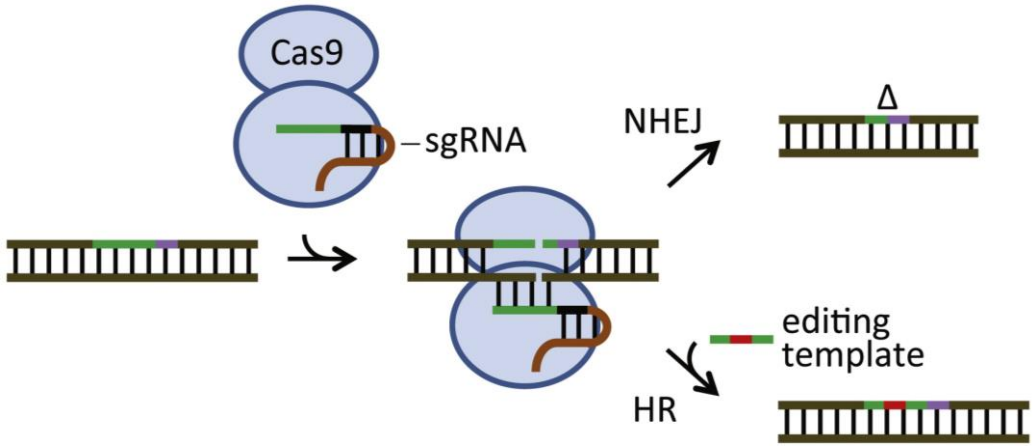


Figure 2 – Cas9-(s)gRNA mediated DNA modification.

Cas9-gRNA complex acts on target DNA, which is homologous to the fragment of gRNA (green), introducing a double strand break. This break can be repaired by endogenous DNA repair mechanisms producing the desired DNA modification (Barrangou and Marraffini, 2014). PAM region and tracrRNA are shown in purple and brown respectively.

Type-II CRISPR-Cas system derived from the bacteria *Streptococcus pyogenes* is the most extensively used RGN in gene editing for several reasons (Doudna and Charpentier, 2014; Hsu et al., 2014; Sander and Joung, 2014; Selle and Barrangou, 2015). First, the RGN in *S. pyogenes* uses Cas9 nuclease, which is a single multi-unit protein and

easy to work with in comparison to other RGNs which need multiple Cas proteins for the proper activity. Second, Cas9 nuclease introduces a single double strand break instead of progressively shredding the target DNA, as for example, in the case of Cas3. Third, the CRISPR-Cas9 system works with crRNA and tracrRNA which are fused into a single transcript and work as a single guide RNA (gRNA) or guide RNA (gRNA) (Jinek et al., 2012). Engineering and delivering a single RNA transcript is more efficient *in vivo*. Chimeric gRNA has been engineered to direct the Cas9 nuclease to cleave complementary DNA when followed by a 5'-NGG PAM in eukaryotic cells (Mali et al., 2013a; Cong et al., 2013; Mali et al., 2013b). Thus, a 5' NGG followed by a 20-nucleotide gRNA attached to the Cas9 nuclease at 3' end can be used to introduce a double strand break at N_{20}-NGG locus in a genome. The only limitation to this is that the PAM needs to be adjacent to the target DNA.

In the past several years, Cas9-gRNA system has been used extensively for genome editing in a large variety of species and cell types including rice, wheat, Arabidopsis, tobacco, sorghum, *C. elegans*, fruit flies, rats, mice, salamanders, frogs, monkeys and human and mouse tissue culture cells (Terns and Terns, 2014). Their therapeutic implications have been demonstrated in targeting antibiotic-resistant and highly virulent strains of bacteria (Bikard et al., 2014; Citorik ae al., 2014), treating HIV (Hu et al., 2014; Ye et al., 2014) and hepatitis B (Zhen et al., 2015) and somatic gene therapy (Ran et al., 2015). Gene editing in primate embryos (Niu et al., 2014) has also been demonstrated while the idea of editing in human embryos for prevention of non-complex hereditary diseases has triggered extensive ethical discussion extending to alteration of complex traits for a benefit (Baltimore et al., 2015).

1.1.3 CRISPR-Cas systems have off-target activity

Cas9-gRNA system uses short RNA sequences to introduce a double strand break in target DNA followed by a PAM sequence, that can be exploited by endogenous DNA repair mechanisms for the desired effect. The targeting is achieved by designing a ~20-nucleotide sequence of gRNA complementary to the target. However, other loci in the genome may also be cleaved if they have sufficient homology to the target DNA, enough for the gRNA to match non-specifically. This potential off-target activity by engineered nucleases is a major concern for their adverse effects in therapeutic applications.

This lack of specificity has been demonstrated in ZFNs (Pattanayak et al., 2011; Gabriel et al., 2011), which may cleave other similar sequences along with their intended targets. Studies have shown that this off-target cleavage activity often introduces undesirable mutations leading to adverse and/or unpredictable effects (Cornu and Cathomen, 2010; Ramirez et al., 2012). TALENS have been reported for their off-target cleavage as well, even though they have been suggested to have better specificity in comparison to ZFNs (Tesson et al., 2011; Hockemeyer et al., 2011; Mussolino et al., 2011). In comparison to ZFNs and TALENs, Cas9-gRNA systems have been suggested to have a higher potential for off-target cleavage. There are several rational explanations for this: 1) previous studies have demonstrated that positions away from the PAM region show more freedom and flexibility for gRNA-target DNA binding (Cong et al., 2013; Gasiunas et al., 2012; Jinek et al., 2013; Jiang et al., 2013), 2) RNA-DNA binding in Cas9-gRNA system opens the possibility to non-Watson-Crick base pairing which can enhance off-target binding (Jiang et al., 2013) and 3) the DNA target sequence in Cas9-gRNA system is

relatively shorter, ~20-nt as opposed to ≥36-nt for TALENS, offering a higher probability of an off-target match in larger genomes, such as in mammals.

The binding of gRNA to an off-target DNA is possible when the on- and off-target DNA sequences partially match with each other. The differences between on- and off-target DNA sequence can be grouped into three cases: a) when both are the same length but with base mismatches, b) when off-target site has insertions with one or more extra bases than the on-target DNA and c) when off-target site has deletions with one or more fewer bases than the on-target DNA. Several recent studies have demonstrated that Cas9-gRNA system generate off-target mutations in mammalian cells with considerably high frequency owing to a non-specific cleavage of off-target DNA with base-pair mismatches (case a) to the target DNA (Fu et al., 2013; Hsu et al., 2013; Pattanayak et al., 2013; Cradick et al., 2013; Cho et al., 2014). As an example, Cradick et al. in 2013 demonstrated that similarity in the gene sequence with base-pair mismatches led to a Cas9 cleavage deleting a 7Kb sequence between two cleavage sites in *HBB* and *HBD*. This provides proof for the possibility of high-frequency gross chromosomal deletions in cases of an off-target match. Mismatches in the PAM sequence, however, are less tolerated, although other studies have demonstrated that Cas9 also recognizes an alternative NAG PAM with low frequency (Hsu et al., 2013; Jiang et al., 2013). These results indicate that the off-target activity of Cas9-gRNA systems may limit its applications mainly in large genomes containing multiple DNA sequences differing by only a few mismatches.

1.1.4 Automated Cas9-gRNA activity estimation through quantification of insertion-deletion mutations

This work is centered on investigation of the above-mentioned cases of partial matches between gRNA and off-target DNA in human cells by automated quantification of the Cas9 nuclease activity with the use of Next-generation sequencing and bioinformatics. To understand the effect of mismatches on potential Cas9-gRNA off-target cleavage in human cells, our collaborators performed the following experimental manipulations: gRNA was varied at different positions throughout the guide sequence to mimic insertions or deletions between off-target sequences and RNA guide strand. The cleavage activity of RNA-guided Cas9 at endogenous loci in HEK293T cells transfected with plasmids encoding Cas9 and gRNA variants were catalogued as the mutation rates induced by Non-Homologous End Joining (NHEJ). These samples were then sequenced on a next-generation sequencing platform to record all indel mutations generated during NHEJ repair.

In order to quantify the cleavage activity generated in the form of indel mutations, we developed an automated bioinformatic pipeline with the ability to process large number of gRNA/DNA variations in sequenced samples. We found that the target cleavage while using gRNA variants introduced more and longer deletions than the original gRNA. We also found that gRNA could cause unintended off-target cleavage in a higher incidence than the on-target sites under specific conditions. Our results clearly indicate the need to scan the genome for potential off-target genomic sites while designing RNA guide strands for targeting specific genomic loci.

The open-source pipeline used for automated quantification of CRISPR-Cas activity (in the form of indel mutations) is available at: https://github.com/piyuranjan/NucleaseIndelActivityScript

1.2 Materials and Methods

Deep sequencing to determine activities at genomic loci. Experimental methods described in this section were performed by our collaborators and are covered in Lin et al., 2014. Genomic DNAs from mock and nuclease-treated cells were used as templates for the first round of PCR using locus-specific primers that contained overhang adapter sequences to be used in the second PCR. PCR reactions for each locus were performed independently for eight touchdown cycles in which annealing temperature was lowered by 1°C each cycle from 65 to 57°C, followed by 35 cycles with annealing temperature at 57°C. PCR products were purified using Agencourt AmPure XP (Beckman Coulter) following manufacturer's protocol. The second PCR amplification was performed for each individual amplicon from first PCR using primers containing the adapter sequences from the first PCR, P5/P7 adapters and sample barcodes in the reverse primers. PCR products were purified as in first PCR, pooled in an equimolar ratio, and subjected to 2 × 250 paired-end sequencing with an Illumina MiSeq.

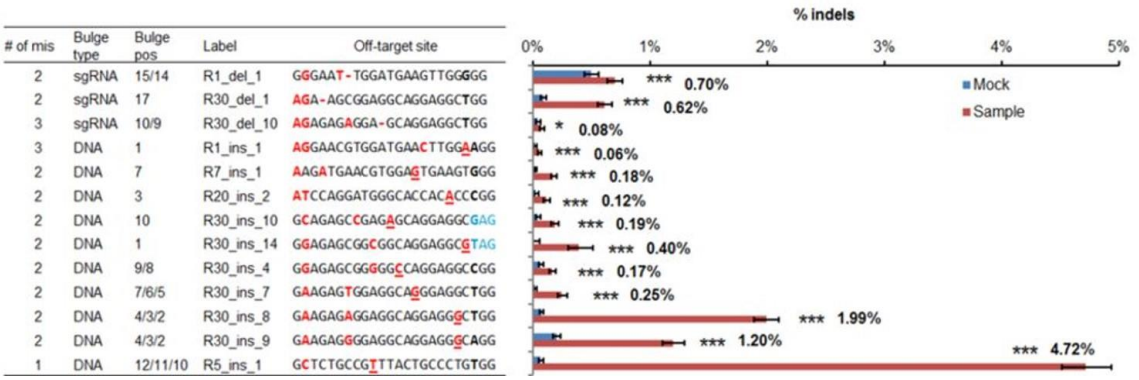
Bioinformatic detection and quantification of indel mutations as a marker for Cas9 activity. Steps describes in this section were followed to analyze indel data and design the automated bioinformatic process. Paired-end reads from the Illumina MiSeq instrument were trimmed from the 3’ end using a trimming-length cutoff with PrinSeq (Schmieder and Edwards, 2011). During this process, the length of the paired-end read is shortened by the

provided cutoff, enhancing efficiency of 3’ adapter removal, in case, the read length extends over the adapters on the 3’ side. The trimming-length cutoff was estimated by subtracting the minimum amplicon size from the average read length (250 in this case) in each sample. A second round of trimming and filtering was performed using TrimGalore (Babraham Bioinformatics), where the read sequences were scanned for the presence of 5’ and 3’ adapter sequences, trimmed to remove any adapter contamination and filtered by a Phred quality score (Q) greater than 20. Paired-end reads were then merged into longer single-end transcripts using FLASH (Magoc and Salzberg, 2011) with a minimum overlap of 10 nucleotides. This merging of the two pairs enhances accuracy for a base call and reduces the chances of an indel call as an artifact of the sequencing platform. Merged reads for each experiment were then aligned to their corresponding reference sequence of the targeted human HEK293T cell line DNA fragment using Burrows-Wheeler Aligner (BWA) (Li and Durbin, 2009) and sorted using SAMtools (Li et al., 2009). Custom alignments spanning ±10-nucleotide of the cleavage site were constructed using SAMtools. Identification and quantification of indel mutations within ±10-nucleotide of the cleavage site was performed with a custom script in Perl. This analysis procedure was later automated in Perl and can be accessed on GitHub (https://github.com/piyuranjan/NucleaseIndelActivityScript/blob/master/indelQuantificationFromFastqPaired-1.0.1.pl). Error bounds for indel percentages are Wilson score intervals calculated using binom package for R statistical software (version 3.0.3) with a confidence level of 95% (R Core Team, 2013). To determine if each off-target indel percentage from a CRISPR-treated sample is significant compared to a mock-treated sample, a two-tailed P-value was calculated using Fisher's exact test.

1.3 Results and Discussion

Deep sequencing was performed at 55 putative off-target sites corresponding to single-base gRNA bulges and 21 sites corresponding to single-base DNA bulges. The sites were amplified from genomic DNA harvested from HEK 293T cells transfected with Cas9 and gRNAs. Putative bulge-forming loci containing one to three PAM-distal mismatches were chosen, since we did not find sites associated with a bulge without any base mismatch. Some of the bulge-forming sites with a high level of sequence similarity, but containing an alternative NAG-PAM were also selected. For comparison, the deep sequencing also investigated 16 on-target sites of the gRNAs tested. Each locus was sequenced from mock-transfected cells as control.

Additional 13 bulge-forming off-target sites with significant cleavage activities resulting from Cas9-gRNA systems compared to the mock-transfected samples were also identified (Figure 3). We found that the number of genomic off-target cleavage sites associated with gRNA bulges was relatively small (some of these cases are indistinguishable from a few mismatches at 5′ end), but there was considerable activity at genomic sites with DNA bulges coupled with one to three additional base mismatches, even with an alternative NAG-PAM. Similar results showing more off-target effect with DNA bulges plus mismatches compared to gRNA bulges plus mismatches were observed in the preliminary T7E1 assay (Lin et al., 2014). The positions of these tolerated DNA bulges are 1-3 and 7-10 bp from PAM, consistent with the results from the model systems using gRNA variants. The majority of the sites with off-target activities detected, as shown in Figure 3, are associated with the gRNA R-30, which has a high GC content (70%). Other gRNAs that resulted in off-target cleavage at bulge-forming loci have GC content ≥50%.



# of mis	Bulge type	Bulge pos	Label	Off-target site
2	sgRNA	15/14	R1_del_1	GGGAAT-TGGATGAAGTTGGGGG
2	sgRNA	17	R30_del_1	AGA-AGCGGAGGCAGGAGGCTGG
3	sgRNA	10/9	R30_del_10	AGAGAGAGGA-GCAGGAGGCTGG
3	DNA	1	R1_ins_1	AGGAACGTGGATGAACTTGGAAGG
2	DNA	7	R7_ins_1	AAGATGAACGTGGAGTGAAGTGGG
2	DNA	3	R20_ins_2	ATCCAGGATGGGCACCACACCCGG
2	DNA	10	R30_ins_10	GCAGAGCCGAGAGCAGGAGGCGAG
2	DNA	1	R30_ins_14	GGAGAGCGGCGGCAGGAGGCGTAG
2	DNA	9/8	R30_ins_4	GGAGAGCGGGGGCCAGGAGGCCGG
2	DNA	7/6/5	R30_ins_7	GAAGAGTGGAGGCAGGGAGGCTGG
2	DNA	4/3/2	R30_ins_8	GAAGAGAGGAGGCAGGAGGGCTGG
2	DNA	4/3/2	R30_ins_9	GAAGAGGGGAGGCAGGAGGGCAGG
1	DNA	12/11/10	R5_ins_1	GCTCTGCCGTTTACTGCCCTGTGG

Figure 3 – Activities of Cas9-gRNA at genomic target and off-target sites with DNA bulges and mismatches.

*Significant activities analyzed by deep sequencing at genomic off-target loci containing bulges coupled with mismatches and in some cases alternative NAG-PAM. Only bulge-containing off-target loci determined to have P-values less than 0.05 are shown. Table on the left shows numbers of mismatches at off-target loci in addition to bulge (no. of mis), bulge types, positions of bulges from PAM (bulge pos), labels for the loci and sequences of off-target sites including PAMs. In these off-target genomic sequences, mismatches are marked by red, deleted base compared to gRNA marked as '–' (gRNA bulge), inserted base compared to gRNA marked as underlined red letters (DNA bulge), NAG-PAMs are marked by blue. Bar graph on the right indicates indel percentages quantified for mock (blue) and treated samples (red) with gRNAs at off-target loci shown in the table to the left. Error bars, Wilson intervals (see 'Materials and Methods' section). *P ≤ 0.05, ***P ≤ 0.001 as determined by Fisher's exact test. The % indel values of treated samples are also indicated.*

Although CRISPR/Cas9 systems can efficiently induce gene modification in many organisms, recent studies revealed that off-target cleavage may occur in mammalian cells with up to five-base mismatches between the short ~20-nucleotide guide RNA and DNA sequences (Fu et al., 2013; Hsu et al., 2013; Pattanayak et al., 2013; Cradick et al., 2013). Research suggests that CRISPR/Cas9 systems can have off-target cleavage when DNA sequences have an extra base (DNA bulge) or a missing base (gRNA bulge) at various locations compared with the corresponding RNA guide strand. Importantly, gRNA bulges of up to 4-bp could be tolerated by CRISPR/Cas9 systems (Lin et al., 2014). The correlation between cleavage activity and the position of DNA bulge or gRNA bulge relative to the

PAM appears to be loci and sequence dependent when comparing the sequence profiles of guide sequences R-01 and R-30. We believe that high GC-content, which makes the RNA/DNA hybrids more stable (Sugimoto et al., 1995), may be responsible for increased tolerance of DNA bulges and gRNA bulges. Guide sequences showing significant bulge-related off-target activity in Figure 3 have high GC content, R-01 50% and R-30 70% with higher representation from R-30, suggesting the importance of high GC content, even with up to three base mismatches.

An interesting finding from this study is that gRNA variants with bulges had different indel spectra than gRNA without bulges. We quantified indel spectra for original gRNAs R-01 (Figure 4) and R-30 (Figure 5), as well as gRNA variants R1 −7/6, R1 C+12, R30 −11 and R30 U+12, using deep sequencing with around 10^4 reads for each sample. Bulge-forming gRNA variants showed higher ratios of larger deletions (Δ10 or Δ7), whereas the original gRNAs without bulges generate mostly 1-bp insertions. This effect is more prominent for variants forming gRNA bulges (R1 C+12 and R30 U+12). Bulge-forming gRNA variants may be more effective than regular gRNAs in creating larger deletions that might be preferred in certain applications, such as targeted disruption of genomic elements.

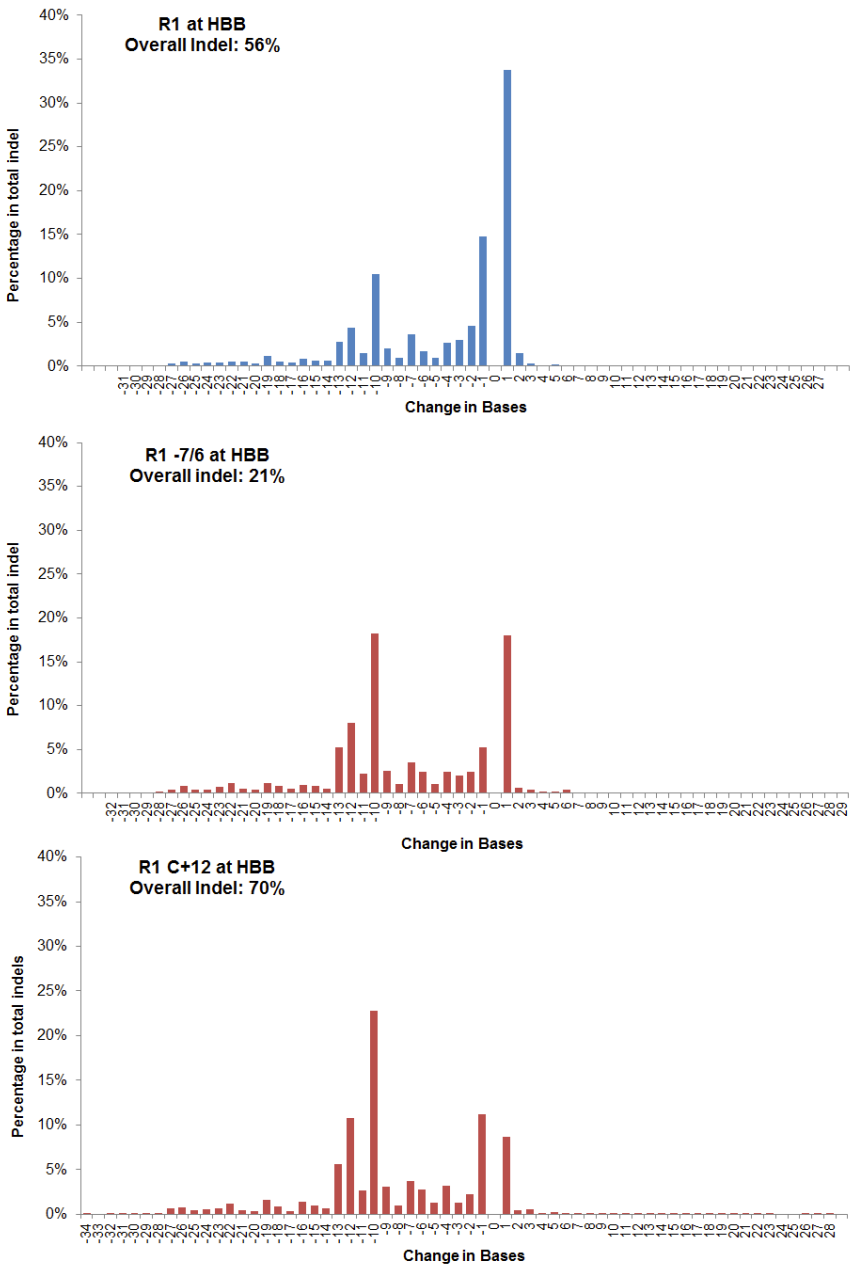


Figure 4 – Indel spectra for original R-01 gRNA and its variants.

Proportional frequency of mutations at different lengths (x-axis) of insertions (positive values) and deletions (negative values) for original sequence of R-01 gRNA (R1, top) and variants for DNA bulge (single base gRNA deletion, R1 -7/6, middle) and gRNA bulge (single base gDNA insertion, R1 C+12, bottom) show a shift from small target DNA insertions (+1) to large deletions (-10). The change in bases at predicted cut sites resulting from R-01 gRNA was calculated using deep sequencing (~10^4 reads per sample). The y-axis represents percentages in all indel reads for R-01 gRNA. Overall % indel in total reads are indicated in each graph.

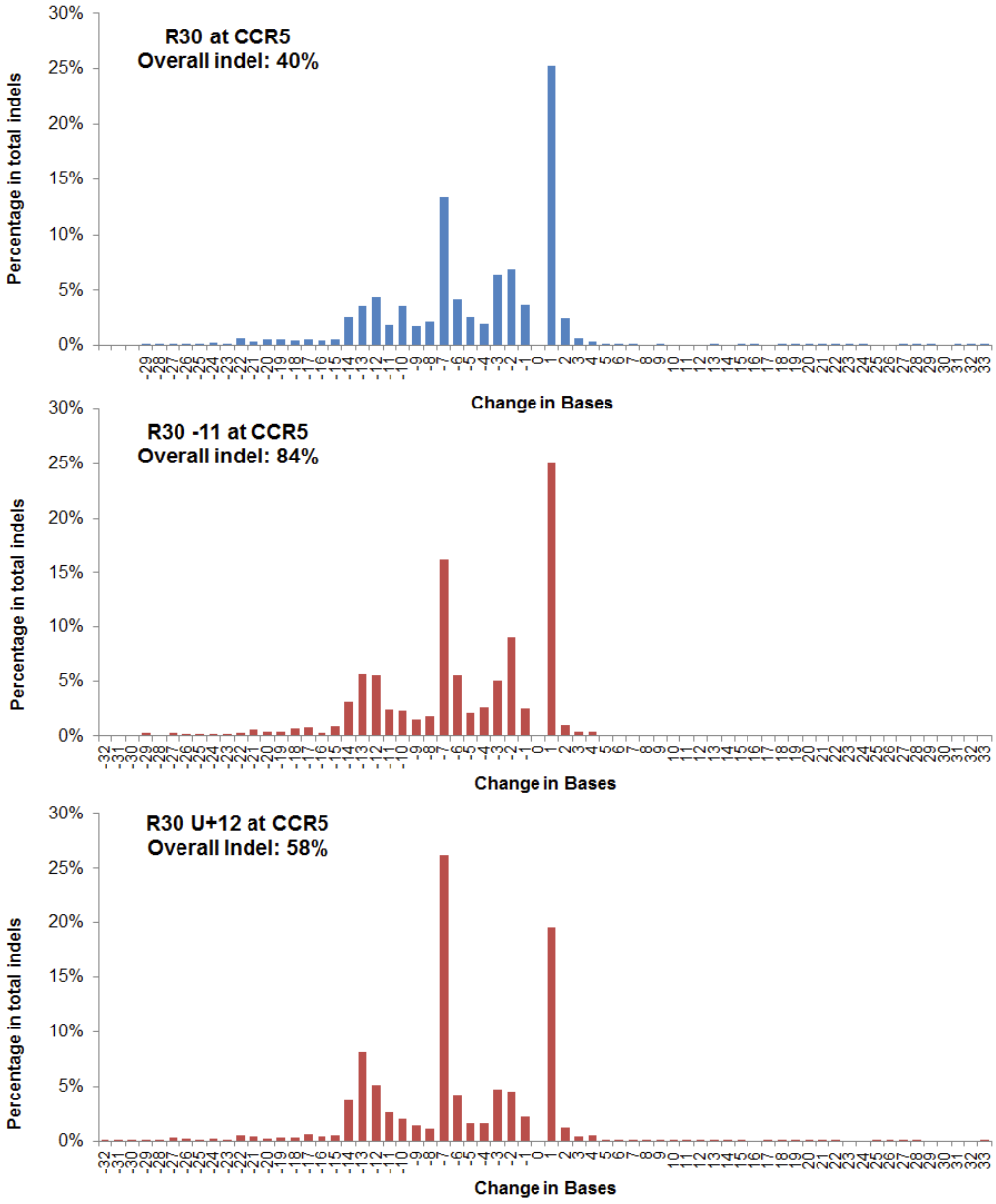


Figure 5 – Indel spectra for original R-30 gRNA and its variants.

Proportional frequency of mutations at different lengths (x-axis) of insertions (positive values) and deletions (negative values) for original sequence of R-30 gRNA (R30, top) and variants for DNA bulge (single base gRNA deletion, R30 -11, middle) and gRNA bulge (single base gDNA insertion, R30 U+12, bottom) show a shift from small target DNA insertions (+1) to large deletions (-7). The change in bases at predicted cut sites resulting from R-30 gRNA was calculated using deep sequencing (~10^4 reads per sample). The y-axis represents percentages in all indel reads for R-30 gRNA. Overall % indel in total reads are indicated in each graph.

1.4 Conclusions

The goal of this research was to investigate the off-target activity of CRISPR/Cas systems owing to partial sequence matching between gRNA and off-target DNA using Next-generation sequencing and automated bioinformatics methods. In order to achieve this goal, gRNA varied at different positions in its sequence was used along with Cas9 to target different combinations of on- and off-target DNA in human HEK293T cell line. The cleavage activity, in the form of indel mutations, was sequenced via next-generation sequencing and analyzed via bioinformatic approaches. Through this research, an automated pipeline method for quantification of CRISPR/Cas activity was developed/. This pipeline can process numerous next-generation sequencing datasets along with related metadata to report Cas nuclease indel activity on DNA targets. This method also generates custom short alignments for the specific loci of the nuclease activity demonstrating all NHEJ repairs performed by the host cell.

Our results suggest the need to perform comprehensive off-target analysis by considering cleavage due to DNA and gRNA bulges in addition to base mismatches. We believe that the following design guidelines will help reduce potential off-target effects of CRISPR/Cas9 systems: (i) conservatively choose target sequences with relatively low GC contents (e.g. ≤35%), (ii) avoid target sequences (with either NGG- and NAG-PAM) with ≤3 mismatches that form DNA bulges at 5′ end, 3′ ends or around 7–10 bp from PAM and (iii) if possible, avoid potential gRNA bulges further than 12 bp from PAM. To aid the rational design of gRNAs for an intended DNA cleavage site, as well as experimental determination of off-target activity, a robust bioinformatic tool that incorporates these design guidelines and ranking potential off-target sites is desired, and more extensive

studies of off-target cleavage by CRISPR/Cas9 systems may be needed concerning the dependence of off-target activity on the type (base mismatch, DNA bulge, gRNA bulge), location and length of sequence differences.

CHAPTER 2. ANALYZING CRISPR-CAS SYSTEM ELEMENTS TO UNDERSTAND VIRAL-MICROBE DYNAMICS IN OMZ COMMUNITIES

2.1 Introduction

2.1.1 Oxygen Minimum Zones in the global ocean

Oxygen concentration in the marine environments is a key factor in balancing marine ecosystems. While it greatly influences the macro- and micro-ecology of the marine ecosystems by structuring the temporal and spatial distribution of all marine life forms, it is, in turn, influenced by physical processes like circulation, salinity and temperature as well as biological factors like microbial production and respiration (Paulmier and Ruiz-Pino, 2008). Most ocean waters are well ventilated through global marine-water circulation patterns and thus maintain a healthy dissolved oxygen concentration throughout the water depths. However, certain locations can see a deficiency of dissolved oxygen in mid-water depths owing to a combination of physical and biological processes like ocean upwelling, high surface productivity, sluggish circulation and stagnation of water masses in different depths due to temperature. Such bodies of water with an oxygen concentration below 20µM are generally referred to as the oxygen minimum zones (OMZs) (Ulloa and Pantoja, 2009).

There are three permanent and intense OMZs each covering a huge geographical expanse in the eastern tropical North (ETNP) and South (ETSP) Pacific Ocean and the Arabian Sea (Paulmier and Ruiz-Pino, 2008; Thamdrup et al., 2012, Karstensen et al., 2008; Ulloa and Pantoja, 2009). Other open ocean OMZs include the northeast sub-arctic

Pacific (NESAP) and regions of the Northwest-African upwelling and the Namibian or Benguela upwelling. Apart from these, there are several other seasonal, enclosed or semi-enclosed OMZs in the Baltic Sea (Conley et al., 2002), Black Sea (Jorgensen, 1982), Cariaco Basin (Scranton et al., 2001), and Saanich Inlet (Anderson and Devol, 1973).

Permanent OMZs are primarily formed along the western boundaries of continents and around the equator where ocean upwelling brings nutrient rich water from the deeper depths to the surface, more so, near the coastal margins. This nutrient availability coupled with the abundance of sunlight fuels extensive primary productivity in the photic zone. The abundance of phytoplankton, which primarily photosynthesize, increase oxygen concentration in the surface waters. Thus, oxygen availability and abundance of organic material, when these phytoplankton die and sink, promote microbial communities that respire and deplete oxygen concentration in the lower depths quicker than it can get resupplied from the surface waters (Wyrtki, 1962). This phenomenon coupled with 1) water stagnation due to warm-to-cold temperature gradient from surface ocean to deeper depths and 2) poor intermixing of the ocean water between the surface and deeper depths lead to over-utilization of oxygen and its subsequent depletion in the mid-water depths (Stramma et al., 2008). The high demand and low availability pulls the oxygen below detection levels at which point the water forms an anoxic marine zone (AMZ) (Ulloa et al., 2012). Oxygen deficiency associated with OMZs stress mobile macro-organisms while the constant sinking of organic rich particles creates a niche suitable for anaerobic microbial processes (Canfield et al., 2010; Stewart et al., 2012; Thamdrup et al., 2012).

2.1.2 Microbial community in the oxygen minimum zones

Bacteria and Archaea demonstrate a highly diverse genetic repertoire including the use of alternate electron donors and acceptors during energy metabolism. These genomic capabilities allow them to utilize niches with low oxygen concentrations helping them in becoming the dominant members of the communities in these ecosystems. Thus, communities along the oxycline and through the core of the OMZ are phylogenetically and metabolically diverse relative to the rest of the ocean (Bryant et al., 2012). Communities in the oxycline contain microaerophilic assemblages, which include nitrite- and ammonia-oxidizing microbes, along with microbes capable of anaerobic metabolism. Communities in the OMZ core, on the other hand, are primarily dominated by microbes capable of anaerobic autotrophic and heterotrophic processes that utilize nitrogen, sulfur and carbon compounds as electron donors and terminal acceptors (Stewart et al., 2012; Padilla et al., 2017; Ulloa and Pantoja, 2009; Ulloa et al., 2012; Wright et al., 2012; Stevens and Ulloa, 2008; Lavik et al., 2009; Walsh et al., 2009; Canfield et al., 2010).

In non-OMZ systems, microbial community distribution, composition and activity has been demonstrated to be influenced by the presence of aggregate particles and microbial adherence to them (Eloe et al., 2010; DeLong et al., 1993; Hollibaugh et al., 2000; LaMontagne and Holden, 2003). Typically, a prefiltering step is used in order to differentially capture microbial communities according to their particle size. To separate communities adhering to aggregate particles from the freely floating community, two collection filters of pore sizes 0.8-30 and 0.2 μm are used subsequently in the biomass filtration and collection line. The microfilter fraction (cell sizes between 0.2 and 1.6–3 μm) is presumed to capture free-living (FL) non-surface attached microbes (Cho and Azam,

1998). The prefilter fraction (cell sizes over 0.8 µm), however, captures particulate aggregates composed of organic debris and surface attached microbial cells (marine snow) along with larger free-living microbes and zooplankton. These particulate aggregates present unique microhabitats with potentially steep gradient of redox substrates along with an abundance of nutrients (Stocker, 2012; Alldredge and Cohen, 1987; Alldredge and Silver, 1988). The free-living microbes on the other hand, experience a comparatively lower nutrient availability and a more stable redox concentration around them. This difference forces particle-associated communities to differ significantly than the free-living communities in composition and their phylogenetic distribution (Hunt et al., 2008; Grossart et al., 2006; Kellogg and Deming, 2009; Plough et al., 2009; DeLong et al., 1993).

2.1.3 Viral-microbial community interaction in the oxygen minimum zones

Viruses play an important role in shaping marine ecosystems and represent an abundant component (Suttle, 2005) with 10^6-10^8 viruses per ml of marine water, generally in higher concentrations in near-shore regions than the open ocean (Hara et al., 1991; Fuhrman and Suttle, 1993; Parson et al., 2011; Proctor and Fuhrman, 1990;). Most of marine viruses are phages that infect autotrophic and heterotrophic bacteria and archaea. Due to this, viruses act as significant agents of microbial mortality, shaping their diversity and the whole marine ecosystem in the process influencing nutrient transformation and cycling (Suttle, 2005; Suttle, 2007). This predator-prey interaction between viruses and microbes is estimated by the viral load in a community or, more precisely, by the viral-to-microbial cell ratio (VMR) (Clasen et al., 2008; Weinbauer et al., 1993; De Corte et al., 2012). Studies have reported VMR in the range of 3-50 for productive surface oceanic waters with a trend of increasing values in the deeper depths, for example, in the Atlantic

waters (Hara et al., 1996; Wommack and Colwell, 2000). Bodies of water with lower oxygen concentrations demonstrate a similar range albeit having a huge variation. However, Cassman et al. in 2012 reported that the permanent anoxic OMZ in the eastern tropical south Pacific Ocean (ETSP) exhibits a lower average VMR in all water layers in comparison to all other marine regions. Their study suggests that the pattern of variation in VMR over depth indicates an oxygen concentration dependency for the influence of viruses on the microbial community and vice versa. Their findings indicate that the ETSP OMZ virome characteristics changed across the depth and oxygen gradient. They demonstrated that the ETSP OMZ harbors novel viruses that showed little genotypic similarity to existing genomes or metagenomes from other marine environments as well as minimal genetic overlap between depths in the water column spanning the surface, oxycline and the anoxic core.

Viruses infect the microbial colonies that may either lead to colony extinction or affect the life cycle of the host through incorporation of its own genetic material in the host genome (Ptashne, 2004). To resist this pressure, bacteria and archaea have developed various types of defense mechanisms, for example, restriction-modification systems which provide them a better chance of survival (Arber, 1979). Clustered regularly interspaced palindromic repeats (CRISPRs), which are a part of a CRISPR-Cas system, are one such way to protect the host genome from invading viruses or other foreign genetic material like plasmids. CRISPR-Cas system works in three stages where 1) Cas proteins cleave a small section of the invading genome and insert it as a spacer between palindromic repeat units in a CRISPR array, 2) the whole CRISPR array is then transcribed and matured by other Cas proteins to form a complex with CRISPR-RNA (crRNA) and a Cas nuclease and

finally, 3) crRNA-Cas complex can act on the invading genome where crRNA can bind to the invader genome via homology and Cas nuclease cleaves or shreds the genome. Apart from imparting immediate defense, this system has other benefits as well. First, because of its property to modify the host genome, this system is heritable to all further generations of the host, giving all future generations an immunity to the invaders the host has encountered. Second, since spacer insertion takes place on one end of the CRISPR array, it could be used to obtain a chronological record of the immunity the host has acquired.

Dynamic interaction between viruses and microorganisms in natural environments is of particular interest (Andersson and Banfield, 2008; DeLong et al., 2006). Studies in the past have investigated virus-host interactions involving CRISPRs within selected bacterial or archaeal species but these interactions are better understood at population and community levels that can provide insights into the relationship between such interactions and host population diversity and density (Sorokin et al., 2010). Metagenomics, in this context, serves as an excellent technique that can use sequenced DNA fragments from a location to represent the microbial community residing in an ecological niche. To develop a better understanding of this relationship in marine environments, Sorokin et al., using metagenomic datasets from Sorcerer II Global Ocean Sampling (GOS) expedition, demonstrated that CRISPR arrays in bacteria and archaea retain the memory of the local viral population making CRISPRs an excellent study system in understanding the forces that shape evolution and diversity of host communities. Oxygen minimum zones, however, are large marine features that are host to unique microbial communities owing to their complex nutrient availability substructure (Stewart et al., 2012; Padilla et al., 2017; Ulloa and Pantoja, 2009; Ulloa et al., 2012; Wright et al., 2012; Stevens and Ulloa, 2008; Lavik

et al., 2009; Walsh et al., 2009; Canfield et al., 2010). Studies have shown the viral community to shift in abundance and diversity along with the oxygen concentrations in one of the largest oxygen minimum zones, in the Eastern South Pacific Ocean (Cassman et al., 2012). However, this interaction may differ largely between communities living a free-floating planktonic lifestyle (free-living, FL) versus those in association with organic-rich marine particles (particle-associates, PA), owing to the tight spatial organization between microbes, and potentially microbes and viruses, on the particles.

2.1.4 CRISPR-Cas system as a proxy to understanding microbial-viral interaction

In this study, we investigate this viral-microbe relationship between particle-associated and free-living microbial communities in an OMZ environment in the Eastern North Pacific Ocean (ETNP) by analyzing CRISPR elements embedded in the host genomes. Using metagenomic datasets from multiple mid-water depths spanning the oxygen gradient from two OMZ locations in the ETNP (Station 02 and 06, Figure 6), we characterize the abundance of CRISPR-Cas systems in PA versus FL communities. We found that CRISPR arrays were present at higher abundance in the PA fraction, supporting the idea that PA communities are more likely to be influenced by viral interactions, potentially owing to the tighter spatial structure and proximity to neighboring host cells.

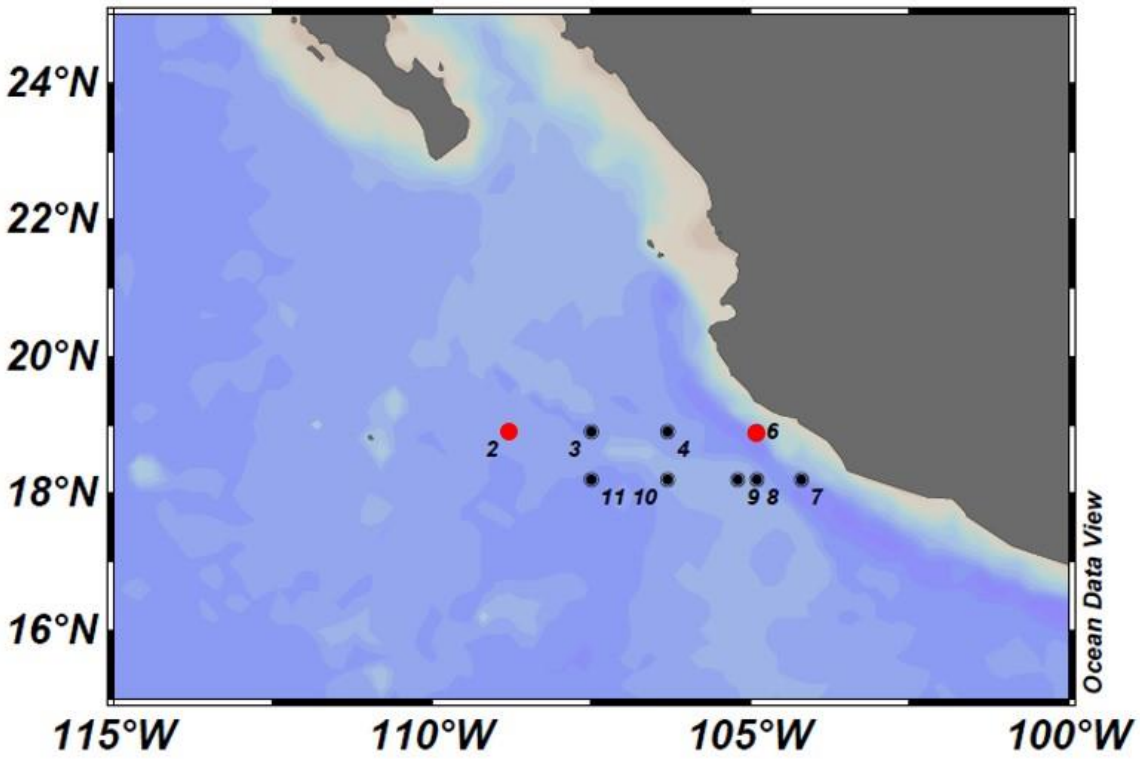


Figure 6 – Geographic locations of the sampling sites in the ETNP OMZ.

All stations marked with numbers were sampled during the ETNP 2013 cruise. Stations marked in red are used in this study. Stations 6 and 2 represent near-shore and off-shore OMZ samples.

2.2 Materials and Methods

Sample collection and metagenomic sequencing. Samples and associated biogeochemical metadata were collected from the ETNP OMZ during the OMZ Microbial Biogeochemistry Expedition Cruise (*R/V New Horizon*, 13-28 June 2013). Subsets of these samples were published in a prior study (Ganesh et al., 2015). Sea water was sampled from different depths spanning the oxygen gradient at different locations including Station 06 (18° 54.0’ N, 104° 54.0’ W) and Station 02 (18° 55.0' N, 108° 48.0' W), which are near-shore and off-shore sites in the ETNP OMZ, respectively, in the same way as described in

Ganesh et al., 2015. For Station 06, the depths sampled were 30 m (upper oxycline), 85 m (lower oxycline), 100 m (secondary chlorophyll maximum), 125 m (secondary nitrite maximum) and 300 m (OMZ core). Briefly, biomass was collected for free-living (cell size 0.2-1.6 µm, 0.2 µm pore size) and particle associated (cell size 1.6-30 µm, 1.6 µm pore size) fractions and was later subjected to whole DNA extraction. Extracted DNA was sequenced according to the standard metagenomic sequencing procedures at the Department of Energy Joint Genome Institute on Illumina HiSeq instrument with 100x2 paired-end reads.

Metagenome assembly, annotation and CRISPR identification. Demultiplexed Illumina paired-end reads were subjected to quality control using a custom wrapper Perl script available at: https://github.com/piyuranjan/MarineMicroLab/blob/master/Scripts/illuminaPeQcTrimMerge.pl. Using this script, 1) paired-end reads were scanned for standard Illumina adapters on 5' and 3' end and trimmed with cutadapt (Martin, 2011) to remove any adapter contamination; 2) read pairs were then quality trimmed with Trim Galore (Babraham Bioinformatics) using a base Phred33 score threshold of Q25 and a minimum length cutoff of 50 bp; 3) high-quality singletons were separated for assembly, while the high-quality paired-end reads were merged with FLASH (Magoc and Salzberg, 2011) by a minimum overlap of 25 nucleotides and 4) merged read sequences (single-end) were exported together with high-quality singletons, produced in a previous step, as one single-end read dataset, while the unmerged (paired-end) were exported as one paired-end read dataset. Both datasets were used as separate libraries for read assembly into contigs using MEGAHIT (Li et al., 2015) assembler, iterating over multiple k-mer values (21, 29, 39,

59, 79, 99, 119, 141), with a minimum multiplicity of 2 and minimum contig length as 200. Basic assembly statistics including N50, L50, largest contig, overall assembly size and unique genes predicted were estimated with QUAST (Gurevich et al., 2013). Features on contigs were predicted through the Prokka (Seeman, 2014) pipeline with RNAmmer (Lagesen et al., 2007) for rRNA, Aragorn (Laslett and Canback, 2004) for tRNA and Prodigal (Hyatt et al., 2010) for protein coding genes. Functional annotation of protein coding genes were added by Prokka through Blast with the default set of core genomes and then by HMM search against a set of default core HMM profiles available in Prokka database. Annotation for CRISPR array with repeat unit and spacer identification was performed using MinCED (derived from CRT, Bland et al., 2007) with repeat length range 19 – 55 and spacer length range 12 – 72, as suggested by previous studies (Anderson et al., 2010; Barrangou and Marraffini, 2014). To determine if the differences in CRISPR array, spacer and Cas protein frequencies between PA and FL communities were significant, a two-tailed P-value was calculated using a paired T-test in R statistical software (R Core Team, 2013).

2.3 Results and Discussion

Metagenomic assembly and annotation of OMZ communities. We assembled community metagenomes each differentiated into two size fractions, from two locations, Station 06 (S06) and 02 (S02) that are a near-shore and an off-shore site, respectively. These metagenomes represent a subset of depths from the surface mixed upper oxycline (30 m in both), oxycline (85 m, 100 m in S06; 80 m, 125 m in S02), OMZ core (300 m) and beneath the OMZ (800 m in S02). The overall assembly size for depths in Station 02 (Table 1) varied between 174.1 – 542.2 Mbp and did not show a significant difference

between the filter fractions except for in the 80 m FL metagenome (1397.5 Mbp) owing to a deeper sequencing of the sample. Overall assembly size for depths in Station 06 (Table 2) varied between 266.2 – 613.2 Mbp and again did not show a significant difference between the filter fractions.

Table 1 – Assembly statistics for ETNP OMZ Station 02 metagenomes

Station 02 MG	Depth (m)	Fraction (PA/FL)	QC Reads ($x10^6$)	Contigs ($x10^3$)	N50 (bp)	L50 (#)	Total length (Mbp)	Largest contig (Kbp)	Unique genes ($x10^3$)	Total genes ($x10^3$)
30PA	30	PA	33.9	240.1	879	68,665	216.1	131.0	354.4	354.6
30FL	30	FL	31.8	534.6	924	153,797	495.5	157.7	852.5	853.0
80PA	80	PA	32.9	203.3	797	58,043	175.3	88.4	254.0	254.0
80FL	80	FL	85.4	135.0	1,064	334,158	1,397.5	1,094.9	2,256.5	2,258.3
125PA	125	PA	38.9	209.6	952	51,827	205.2	178.4	285.5	285.8
125FL	125	FL	33.0	486.2	1,197	105,210	542.2	263.6	853.0	853.8
300PA	300	PA	43.1	330.0	1,218	69,966	371.8	115.4	501.7	502.3
300FL	300	FL	29.5	293.7	1,189	65,985	325.7	81.4	508.0	508.7
800PA	800	PA	36.7	179.3	948	44,689	174.1	87.5	230.3	230.4
800FL	800	FL	32.9	497.0	1,093	116,148	526.3	300.3	829.0	829.7

Table 2 – Assembly statistics for ETNP OMZ Station 06 metagenomes

Station 06 MG	Depth (m)	Fraction (PA/FL)	QC Reads ($x10^6$)	Contigs ($x10^3$)	N50 (bp)	L50 (#)	Total length (Mbp)	Largest contig (Kbp)	Unique genes ($x10^3$)	Total genes ($x10^3$)
30PA	30	PA	32.9	294.8	870	81,770	266.2	106.1	386.5	386.6
30FL	30	FL	31.2	586.1	1,090	143,927	613.2	221.7	986.5	987.3
85PA	85	PA	29.6	402.1	991	101,679	397.2	64.7	538.1	538.4
85FL	85	FL	34.2	432.0	1,406	81,863	531.3	135.6	796.4	797.3
100PA	100	PA	35.7	435.6	1,050	102,094	447.6	77.7	582.3	582.7
100FL	100	FL	34.6	441.6	1,342	85,736	527.3	149.1	791.4	792.3
125PA	125	PA	28.2	363.1	1,079	79,415	384.0	139.8	506.5	506.9
125FL	125	FL	31.6	399.7	1,327	80,592	473.4	141.0	707.1	708.0
300PA	300	PA	34.8	445.9	1,144	95,380	485.8	139.8	662.7	663.5
300FL	300	FL	30.9	345.6	1,211	74,672	388.2	101.1	595.6	596.4

Variation of CRISPR-Cas elements between particle-associated and free-living communities. For consistency, to estimate the relative abundance of CRISPR-Cas system elements (CRISPR arrays, spacers and Cas proteins) in metagenomic contigs, we divided the total number of elements found by the total length of the sequences assembled. In summary, CRISPR elements in the ETNP OMZ microbial communities were differentiated by community lifestyle irrespective of the oxygen concentrations. The abundance of CRISPR arrays and spacers differed between particle-associated and free-living communities consistently in the near-shore (Station 06) OMZ water column, but this pattern is less pronounced in the off-shore (Station 02) water column. To elaborate, CRISPR arrays and spacers were at higher proportional representation in the particle-associated communities in longer contigs (>1Kb) and all assembled contigs in Station 06 in comparison to free-living communities (Figure 7). This difference is statistically significant between both communities in all water depths at Station 06 (Figure 9) as determined by a two-tailed paired T-test. The higher abundance of CRISPR arrays indicates that more host genomes in the community have acquired CRISPR-Cas mediated immunity while a higher frequency of spacers indicate that over time, the host genomes have experienced more viral infections in the particle-associated communities. These patterns suggest that the particle-associated communities observe a higher abundance and proliferation of viral communities suggesting more frequent predator-prey interactions, although this hypothesis remains to be tested. Spatially tighter structuring of PA communities might facilitate these interactions giving viruses more host cells to infect. In contrast, the host availability in FL communities could be much lower.

A similar, statistically significant, trend is observed in all assembled contigs in Station 02 (Figure 8, Figure 9). However, the longer contigs show a marginal, non-significant, increase with the exceptions of 1) communities at 80 m showing a higher rate in the FL fraction, which is higher than any other FL associated rates as well and 2) communities in the OMZ core (300 m) showing a higher rate in PA fraction, which is significantly higher than any other PA associated rates. For *cas* genes, the trend was highly variable and statistically insignificant in both stations.

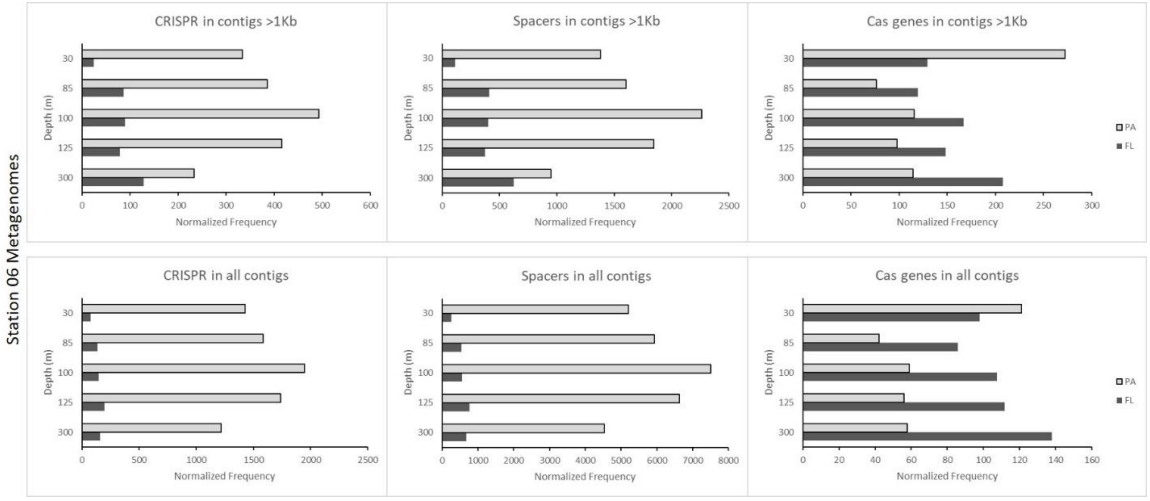


Figure 7 – Frequencies of elements of CRISPR-Cas systems in Station 06 ETNP OMZ metagenomes.

Frequency of CRISPR arrays (left), spacers (middle) and Cas proteins (right) are shown for all depths spanning the oxygen gradient in PA vs FL communities. Top and bottom panels show frequencies in assembled contigs larger than 1Kb and in all assembled contigs. All frequencies are normalized by their corresponding assembly size.

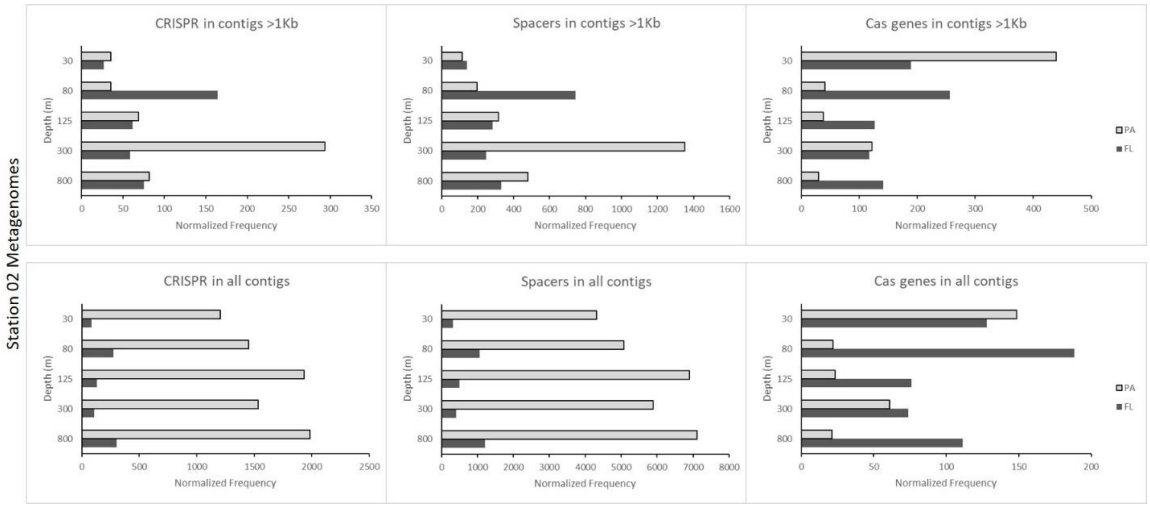


Figure 8 – Frequencies of elements of CRISPR-Cas systems in Station 02 ETNP OMZ metagenomes.

Frequency of CRISPR arrays (left), spacers (middle) and Cas proteins (right) are shown for all depths spanning the oxygen gradient in PA vs FL communities. Top and bottom panels show frequencies in assembled contigs larger than 1Kb and in all assembled contigs. All frequencies are normalized by their corresponding assembly size.

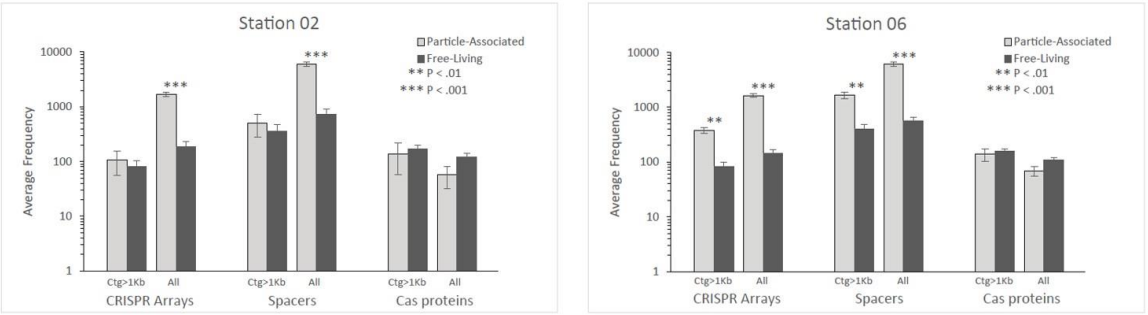


Figure 9 – Comparison of frequencies of CRISPR-Cas system elements between PA and FL communities.

Frequency of CRISPR-Cas elements (CRISPR array, spacers and Cas proteins) in ETNP OMZ Stations 02 (left) and 06 (right) shown are averaged over all depths. Error bars show standard error, with significance (P-value) calculated by 2-tail paired T-tests.

Variation of CRISPR-Cas elements along the oxygen gradient. In summary, CRISPR-Cas characteristics in the ETNP OMZ microbial communities changed across depth and oxygen concentration. Frequency of CRISPR arrays and spacers in the near-

shore (Station 06) OMZ water column appear to coincide with changes in viral abundance and diversity in other OMZ water columns as reported by previous studies (Cassman et al., 2012). However, the pattern at the off-shore site (Station 02) differed from that at station 06 closer to shore. To elaborate, CRISPR arrays and spacers were found in higher abundance in the particle-associated communities in Station 06 (Figure 7) in the oxycline depths (85, 100 m in S06) but gradually declined towards the core OMZ depths (125 m nitrite max, 300 m core) while a more variable trend can be observed in the free-living communities. Assuming that a higher CRISPR and spacer frequency in the host community genome would correlate with a higher abundance and prominence of the viral community, our findings are consistent with previous studies reporting higher-to-lower-to-higher viral abundances while going from oxygenated surface waters to the OMZ core to benthic oxygenated waters (Vik et al., 2017). A similar trend is observed in all assembled contigs in both fractions (PA and FL) at Station 02 (Figure 8) with minima in the core OMZ depth (300 m) and abundances rising again with oxygen concentration in the lower oxycline (800 m). Longer contigs (>1Kb) in the same site show a similar OMZ core minimum but only in the free-living communities, while a variable trend with a peak in identified CRISPRs and spacers in the core OMZ depth can be observed in the particle-associated communities. *cas* genes, however, show a highly variable trend in nearly all samples, including both stations, communities and assembly lengths. Better refining the pattern in *cas* genes can perhaps be accomplished by a more exhaustive search for *cas* genes in these metagenomic assemblies by developing a more comprehensive custom database of Hidden Markov Model (HMM) profiles using all known *cas* genes to date.

2.4 Conclusions

Predator-prey interactions between viruses and microbes play a significant role in shaping marine microbial communities. To defend themselves, bacteria and archaea have developed a guided nuclease strategy via the CRISPR-Cas system that provides them specific immunity against concurrent viral infections and keeps a signature of the infections the host and its predecessors have encountered in the host genome as CRISPR arrays. These CRISPR arrays can be used to understand viral dynamics and prevalence in a microbial community. This study presents the first metagenomic comparison of the elements in the CRISPR-Cas system between microbial communities living a free-floating planktonic lifestyle versus an organic rich particle-attached lifestyle in a marine oxygen minimum zone environment.

In our analysis, we found that abundance of CRISPR arrays and spacer units were higher in the microbial communities living a particle-associated lifestyle versus the free-floating planktonic lifestyle. These results suggest that the particle-associated communities experience more frequent predator-prey interaction possibly owing to host proximity in a spatially tighter microbial community. We also observed that the vertical distribution of CRISPR arrays and spacers coincides with the previous estimates of viral abundance and diversity in OMZ environments (Cassman et al., 2012; Vik et al., 2017). Notably, the oxycline exhibits an increase in the abundance of CRISPR arrays and spacers. However, the OMZ core exhibits a lower abundance of CRISPR arrays and spacers suggesting more limited host-virus dynamics. We do believe that better annotations for *cas* genes and characterization of spacers (viral) and repeat (microbial) elements can improve our understanding of OMZ microbial-viral dynamics.

CHAPTER 3. PHYLOGENETIC RESOLUTION AND CHARACTERIZATION OF JS-1 ATRIBACTERIA IN MARINE METHANE HYDRATES.

3.1 Introduction

The aims of this research were 1) to extract a JS1 type Atribacteria genome from the whole microbial community using genome-directed metagenomic analysis approaches and 2) estimate the phylogenetic diversity within the microbial population constituting the JS1 clade using 16S rRNA targeted gene sequencing. This study was a part of a larger effort to understand niche-specific adaptations in JS1 Atribacteria that help them maintain a high proportional abundance in the microbial communities living in a marine methane hydrate sediment environment. To achieve this goal, our collaborator and lead PI on this project (Dr. Jen Glass) sampled the microbial communities from different depths in the sediment and extracted the whole-community DNA. These communities were sequenced on a next-generation sequencing platform and were analyzed through the metagenomic analysis steps detailed in this chapter.

3.1.1 Atribacteria as a candidate phylum

Atribacteria is a recently proposed candidate phylum in the Bacteria that is globally distributed and identified from geothermal springs and methane rich marine sediments (Nobu et al., 2016). Single-cell amplified genome (SAG), metagenome and 16S rRNA gene sequencing have been instrumental for identifying Atribacteria genomes and their abundance in various anoxic marine and non-marine sedimentary environments (Carr et

al., 2015). Using 16S rRNA gene sequencing, Atribacteria have been identified from mud volcanoes (Niemann et al., 2006), brackish sediments (Webster et al., 2004; Rinke et al., 2013), hydrothermal areas (Teske at al., 2002), tidal flats (Wilms et al., 2006; Webster et al., 2007), organic-rich deep-sea sediments (Webster et al., 2006) and methane hydrate bearing sediments of the Sea of Okhotsk, the Nankai Trough and the Peru and Cascadia margins (Newberry et al., 2004; Inagaki et al., 2003, 2006; Reed et al., 2002). Although, candidate phyla are typically found in low abundance (Sogin et al., 2006; Elshahed et al., 2008), sediments rich in methane hydrates have demonstrated a higher relative abundance of Atribacteria, representing over 50% of the community in many cases (Newberry et al., 2004; Inagaki et al., 2003, 2006), suggesting niche specific adaptations in Atribacteria for these environments.

Depending on the environment (marine/non-marine) from which Atribacteria have been isolated, they have been observed to group into two monophyletic lineages, OP9 and JS1. When originally identified, the JS1 lineage was thought of as a deeply branching lineage under the candidate phylum OP9 (Hugenholtz et al., 1998; Webster et al., 2004), but recent phylogeny estimates have described them as two very-distinct monophyletic lineages grouped together under Atribacteria (Nobu et al., 2016). OP9 was first discovered in the sediments of the hot spring Obsidian Pool in Yellowstone National Park, USA (Hugenholtz et al., 1998). Since then, targeted 16S rRNA gene surveys have demonstrated the presence of OP9 lineage in wastewater digesters, biogas reactors (Leven et al., 2007; Riviere et al., 2009; Wrighton et al., 2008; Tang et al., 2011), petroleum reservoirs (Kobayashi et al., 2012; Gittel et al., 2009) and a variety of terrestrial geothermal springs (Vick et al., 2010; Lau et al., 2009; Wemheuer et al., 2013; Costa et al., 2009; Sayeh et al.,

2010). JS1 on the other hand, has been observed in higher abundances in the marine sub-seafloor environments, particularly from sediments around continental margins and shelves that are rich in organic substrates, for example, hydrocarbon seeps and methane hydrate bearing sediments (Orcutt et al., 2011; Parkes et al., 2014; Inagaki et al., 2006). JS1 sequences have also been found in hypersaline microbial mats (Harris et al., 2013), landfill leachates (Liu et al., 2011) and petroleum reservoirs (Kobayashi et al., 2012; Pham et al., 2009).

3.1.2 Characterization of OP9 and JS1 type Atribacteria

While 16S rRNA sequencing provides information about community abundance, Single amplified genomes (SAGs) and metagenome assembled genomes (MAGs) can make it possible to extract partial, near-complete or complete genomes of Atribacteria, helping elucidate the genome potential of this phylum level candidate. Using these genomes in a comparative framework can provide insights into metabolic and phylogenetic diversity within this poorly understood branch of life. Previously, two genomes, named *Candidatus Caldatribacterium californiense* and *Ca. Caldatribacterium saccharofermentans*, belonging to the OP9 lineage were recovered from terrestrial geothermal spring environments (Dodsworth et al., 2013; Chouari et al., 2005). The first was recovered through a genome reconstruction with a co-assembly of 15 SAGs from Little Hot Creek (Vick et al., 2010), while the second was recovered as a MAG from a thermophilic cellulosic consortium in Great Boiling Spring (Peacock et al., 2013). Both these genomes were hypothesized to be living a heterotrophic lifestyle with saccharolytic fermentation in strictly anaerobic conditions. Another single OP9 SAG was recovered from a terephthalate (TA)-degrading bioreactor that was a part of a larger effort towards single-

cell genome sequencing for a large variety of previously unknown or unclassified taxa from different low-oxygen environments (Rinke et al., 2013). The same study also yielded 13 JS1 SAGs from TA bioreactor, sediments from Etoliko Lagoon, Greece and the anaerobic, sulfidic monimolimnion of meromictic Sakinaw Lake, Canada. Other JS1 SAGs have also been isolated from marine sediments in Aarhus Bay, Denmark (Lloyd et al., 2013). Since, the coverage of JS1 type Atribacteria has been reported from a wide range of organic rich environments, we posit a natural phylogenetic diversity in this clade with niche specific adaptations. Also, while the genomes belonging to the OP9 clade have all been suggested as saccharolytic heterotrophs (Carr et al., 2015), metabolism in the diverse JS1 clades still needs more investigation.

In this study, we aim to investigate the previously unexplored phylogenetic diversity within the JS1 Atribacteria that are found abundant in a methane hydrate bearing marine sediment near the Cascadia convergent margin. Along with understanding the diversity, we also aim to recover a genomic isolate for the JS1 Atribacteria from these sediments enabling further investigation of niche-specific adaptations in Atribacteria from this environment. For these purposes, we used targeted 16S rRNA amplicon and metagenomic analysis of the sequence datasets sampling the whole microbial community from seven marine sub-surface sediment depths, 2 – 68.55 m below seafloor (mbsf), in ODP site 1244 near the Cascadia convergent margin.

3.2 Materials and Methods

Sample collection, DNA extraction and Sequencing. Sample collection and experimental methods described in this section were performed by our collaborators (lab

of Dr. Jennifer Glass, Earth and Atmospheric Sciences, Georgia Institute of Technology). Briefly, sediment core samples were drilled from ODP site 1244 on South Hydrate Ridge (44°35.18'N, 125°7.19'W; 600-800 m water depths) on the Cascadia convergent margin on ODP Leg 204 in 2002. The DNA was extracted, from the following depths in meters below seafloor (mbsf): 1.95-2.25 (C1-H2); 3.45-3.75 (C1-H3); 8.60 (F2-H4); 18.10 (F3-H4); 20.69 (C3-H4); 35.65 (E5-H5) and 68.55 (E10-H5). Microbial community composition was assessed by Illumina sequencing of the V3-V4 region of the 16S rRNA gene. Amplicons were amplified via PCR and sequenced on an Illumina MiSeq across two different runs using a 500-cycle kit with 5% PhiX to increase read diversity. Whole microbial community DNA was extracted and sequenced using a Rapid-Run on an Illumina HiSeq 2500 to obtain 100 bp paired-end reads. 16S rRNA sequences were deposited into NCBI SRA SAMN04214977-04214990 (BCO-DMO project 626690) and the metagenome sequences were deposited into NCBI SRA SAMN04044156-04044143 (BCO-DMO project 626690).

Metagenome assembly binning and annotation. Demultiplexed Illumina reads were mapped to known adapters using Bowtie2 (Langmead and Salzberg, 2012) in local mode to remove any reads with adapter contamination. Read pairs were quality trimmed with Trim Galore (Babraham Bioinformatics) using a base Phred33 score threshold of Q25 and a minimum length cutoff of 80 bp. High quality reads were then assembled into contigs using the SPAdes assembler (Bankevich et al., 2012) with --meta option for assembling metagenomes, iterating over a range of k-mer values (21, 27, 33, 37, 43, 47, 51, 55, 61, 65, 71, 75, 81, 85, 91, 95). Assemblies were assessed with reports generated with QUAST (Gurevich et al., 2013). Features on contigs were predicted through the Prokka pipeline

(Seemann, 2014) with RNAmmer (Lagesen et al., 2007) for rRNA, Aragorn (Laslett and Canback, 2004) for tRNA, Infernal (Nawrocki and Eddy, 2013) and Rfam (Nawrocki et al., 2014) for other non-coding RNA and Prodigal (Hyatt et al., 2010) for protein coding genes. Annotation for protein-coding genes was performed as follows: 1) functional annotation through Blast (Altschul et al., 1990) with the default set of core genomes and then by HMM search (Mistry et al., 2013) against a set of default core HMM profiles available in Prokka, 2) functional annotation through the Blast Descriptor Annotator algorithm in Blast2GO (Conesa et al., 2005) that utilizes Blast against the NCBI nr database, 3) KEGG orthology assignment using GhostKOALA (Kanehisa et al., 2016) and 4) functional annotation through InterProScan (Jones et al., 2014) that performs a cross-reference HMM search across multiple databases finding Pfam families and sequence motifs with close homology.

Metagenome contigs were partitioned through MetaBAT (Kang et al., 2015) into genome bins (MAGs) representing individual species using tetranucleotide frequency and sequencing depth. Sequencing depth was estimated by mapping reads on to the assembled contigs using Bowtie2 and Samtools (Li et al., 2009). Completeness, contamination and strain level heterogeneity was assessed using single copy marker genes in CheckM (Parks et al., 2015). Gene features and their functional annotations for MAGs were extracted from the metagenome for the contigs that belong to the bins. Initial taxonomic affiliation for bins was inferred based on LCA affiliation in MEGAN6 (Huson et al., 2007) and by the top Blast matches to the marker gene *rpoB*.

Phylogeny reconstruction for Atribacteria MAGs. Coding sequences (CDS) from the whole genomes were downloaded from the NCBI representative genomes collection using

NCBI e-utilities. The downloaded set included 405 genomes spanning a wide diversity of lineages in bacteria. A single genome candidate per genus was selected for this purpose with the criteria that it has at least 1000 genes. Sequence duplication (100% identity, unlikely to be biological duplication) within genomes was removed using CD-HIT (Li and Godzik, 2006). Available reference Atribacteria genomes, 24 in total, as either SAGs or a combined assembly of SAGs or MAGs (Table 3), were downloaded and annotated using the Prokka pipeline. A list of 139 core single copy genes (CSCG) as HMM profiles was obtained from Rinke et al. 2013. Representative, reference Atribacteria genomes and the E10H5 Atribacteria MAG were then scanned for the presence of these HMM profiles using HMMer with the recommended score threshold for each profile as provided in Rinke et al. 2013. In a series of manual subsampling steps, 69 CSCG clusters were selected from 220 representative genomes and 20 Atribacteria genomes where 1) 69 clusters were present in only a single copy, 2) all 69 clusters were present in 220 representative genomes, and 3) the minimum number of clusters present in any Atribacteria genome was 6. Five of 25 Atribacteria SAG genomes (Etoliko Lagoon SAG 227, Sakinaw Lake SAG 124 and 125, TA biofilm SAG 231 and 232) were excluded since they covered less than 5 of the 69 marker genes scanned. All 69 CSCG clusters were aligned individually using the L-INS-i mode in MAFFT (Katoh and Standley, 2013). Alignments were then concatenated using a custom script Aln.cat.rb from the Enveomics collection (Rodriguez and Konstantinidis, 2016) with invariable sites removed. Phylogeny reconstruction was performed in RAxML (Stamatakis, 2014) using the GAMMA model of rate heterogeneity, iterating over all models of protein substitution to choose the one with best log likelihood, and 1000 rapid bootstraps using the MRE convergence criterion (50 bootstrap replicates performed),

followed by a thorough Maximum Likelihood (ML) search. The resulting phylogenetic tree was modified for optimal viewing in iTOL (Letunic and Bork, 2016) with a full view including all lineages and a pruned view confirming placement of E10H5-B2 genome bin in the Atribacteria phylogeny in more detail.

Table 3 – Genome statistics for other known Atribacteria SAGs/MAGs.

SAG/MAG ID	Clade	Completeness	Contamination	Strain Heterogenity	Contigs	Genome Size (Kbp)	GC (%)	CDS	Predicted Size (Kbp)
Alaska_North_Slope_MG_bin_34-128	JS1-1	64.41	5.08	0	86	894.9	33.42 ± 2.00	781	1389.4
Sakinaw_Lake_SAG_130	JS1-1	62.28	3.99	80	131	1653.3	34.09 ± 2.49	1596	2654.7
Sakinaw_Lake_SAG_71	JS1-1	58.73	3.39	50	183	2196.9	34.26 ± 2.63	2084	3740.7
Sakinaw_Lake_SAG_co-assembly	JS1-1	56.61	2.12	25	153	2091.4	34.27 ± 2.41	1973	3694.4
Sakinaw_Lake_SAG_136	JS1-1	42.2	0.53	66.67	117	1424.4	34.26 ± 2.15	1336	3375.4
Sakinaw_Lake_SAG_218	JS1-1	37.93	0	0	37	497.8	32.92 ± 2.33	464	1312.4
Sakinaw_Lake_SAG_217	JS1-1	32.46	0	0	27	329.1	33.81 ± 2.40	306	1013.9
Sakinaw_Lake_SAG_216	JS1-1	32.2	0.85	100	78	975.9	34.50 ± 2.37	915	3030.6
Sakinaw_Lake_SAG_219	JS1-1	31.03	0	0	83	939.1	33.99 ± 1.94	875	3026.4
Sakinaw_Lake_JS1_MG_bin	JS1-1	25.49	1.85	0	150	345.7	32.53 ± 2.85	308	1356.1
Aarhus_Bay_SAG_I22	JS1-1	22.41	0	0	291	1040.3	35.03 ± 3.48	976	4642.0
Sakinaw_Lake_SAG_213	JS1-1	17.24	0	0	53	436.8	34.00 ± 2.43	440	2533.8
Sakinaw_Lake_SAG_124	JS1-1	8.62	0	0	20	152.6	32.60 ± 2.46	149	1770.6
Sakinaw_Lake_SAG_215	JS1-1	8.33	0	0	32	505.0	34.34 ± 2.10	473	6062.1
TA_biofilm_JS1_MG_bin	JS1-2	73.16	9.5	44.44	790	2050.9	34.79 ± 3.24	1774	2803.2
TA_biofilm_SAG_231	JS1-2	35.34	0	0	52	1011.3	35.40 ± 2.86	955	2861.6
Aarhus_Bay_SAG_B17	JS1-3	8.33	0	0	246	1130.2	33.78 ± 3.88	1111	13567.6
Etoliko_Lagoon_SAG_227	JS1-3	8.33	0	0	28	321.7	33.75 ± 3.50	297	3861.6
Alaska_North_Slope_MG_bin_34-868	JS1-4	19.22	0	0	56	199.8	33.75 ± 2.52	196	1039.4
TA_biofilm_SAG_167	JS1-4	16.95	0	0	27	498.8	34.22 ± 1.37	454	2942.9
GBS_77CS_MG_bin	OP9	99.44	6.39	28.57	306	2252.4	55.35 ± 1.86	2376	2265.0
GBS_77CS_MG_bin_cov-filtered	OP9	99.44	0	0	153	1795.2	55.58 ± 1.57	1899	1805.3
LHC_SAG_co-assembly	OP9	96.61	0.15	0	470	2080.5	53.66 ± 4.10	2223	2153.5
TA_biofilm_SAG_232	OP9	0	0	0	43	719.6	36.81 ± 2.63	639	0.0

16S rRNA gene amplicon analysis. Demultiplexed amplicon read pairs were trimmed for Illumina adapter contamination and base calls with quality lower than Phred score threshold of Q25 using Trim Galore. Read pairs were discarded if any of the pairs were smaller than 100 bp. Paired reads were merged with FLASH (Magoc and Salzberg, 2011) with a minimum overlap criterion of 25 bp. After this point, sequencing duplicates

were used together to improve sequencing depth. Merged sequences were checked for sequencing chimeras with identify_chimeric_seqs.py script in QIIME v1.9.1 (Caprasso et al., 2010) using SILVA (Quast et al., 2012) SSU database release 128 as the reference. For the purposes of phylogenetic placement and diversity assessment of the Atribacteria population, the analysis was performed in two ways:

1) *16S rRNA phylogenetic analysis*: Chimera-checked sequences from the deepest depth (E10H5, 68.55 mbsf) were clustered at 97% similarity using the open reference OTU picking strategy and taxonomy assigned in QIIME using the SILVA release 128 as the reference. Clusters with only a single representative were ignored for further analysis. A representative sequence from each cluster identified as Atribacteria OP-9 and JS-1 was extracted for alignment. The relative and absolute abundances of these representative sequences in the sample was calculated. Reference Atribacteria 16S rRNA gene sequences were collected from Nobu et al. 2016, Carr et al. 2015 and Yarza et al. 2014. Atribacteria 16S rRNA gene sequences from other already published SAGs and MAGs were extracted from their annotations generated using Prokka. All reference sequences collected (16S clones and SAGs/MAGs), totaling 286, were aligned in MAFFT with the linsi option, alignment reordering and reverse complement matching enabled. Eight Firmicute sequences were added to the alignment as outgroup. After manual inspection and correction, the alignment consisted of 273 sequences and 1485 positions and here onwards is referred to as the reference alignment (RA).

a) *Phylogenetic inference with Atribacteria OTU sequences*: Without modifying the RA positions, 230 Atribacteria representative OTU sequences were added through MAFFT using previously described options. Alignment was then manually inspected and trimmed

to include only the V3-V4 region. After inspection, alignment consisted of 498 sequences and 275 positions. This alignment was then used for phylogeny reconstruction in RAxML with the GTRGAMMA model and 500 rapid bootstraps, followed by a thorough ML search. The resulting phylogenetic tree was modified for optimal viewing in iTOL. Pairwise distances between all Atribacteria sequences and OTUs (excluding Firmicute sequences) in this alignment were calculated using the p-distance method in MEGA7 (Kumar et al., 2016) and summarized in R (R core team, 2013) as: min 0.0, 1st quartile 0.5, median 0.09, mean 0.11, 3rd quartile 0.18 and max 0.27. Pairwise distances between only OTUs (excluding Firmicute and Atribacteria reference sequences) in this alignment were calculated using the p-distance method in MEGA7 and summarized in R as: min 0.004, 1st quartile 0.056, median 0.075, mean 0.076, 3rd quartile 0.095 and max 0.194.

b) *Phylogenetic inference with evolutionary placement of Atribacteria OTU sequences among reference Atribacteria clades*: The reference alignment, RA (1485 positions), was used for phylogeny reconstruction in RAxML with the GTRGAMMA model and 500 rapid bootstraps followed by a thorough ML search. The resulting phylogenetic tree was used as the reference Atribacteria phylogeny which demonstrates the clade structure within Atribacteria lineages. This reference phylogeny along with the previously prepared V3-V4 region alignment for all Atribacteria OTUs (498 sequences, 275 positions) was used in RAxML Evolutionary Placement Algorithm (EPA) for placement of Atribacteria OTUs among the reference Atribacteria clades. RAxML EPA was primed for selection of top 10% branches for ML based thorough insertion, reporting top (up to 7 placements, within 1% of maximum insertion score) insertions for all sequences during this analysis. The resulting placements were modified for optimal view in iTOL.

2) *16S rRNA diversity analysis*: Chimera-checked sequences were grouped together from all depths for diversity analysis, clustered at 97% similarity using the open reference OTU picking strategy and taxonomy assigned in QIIME using SILVA release 128 as the reference database. A core set of QIIME diversity analysis was performed using the core_diversity_analyses.py workflow script at an even sampling depth of 29194 sequences. The phylogenetic diversity (PD) metric (Faith, 1992) was used to quantify alpha diversity across samples. The weighted Unifrac metric (Lozupone & Knight, 2005) was used to calculate beta diversity, and sample relatedness was visualized on a two-dimensional principal coordinate analysis plot.

Gene orthology analysis between B2 genome and other reference Atribacteria: 23 reference Atribacteria (MAGs/SAGs) genomes were annotated using Prokka. The predicted genes in all genomes were used with genes identified in the B2 genome in a Blast best hit (BBH)-based clustering analysis for orthologous group identification through Proteinortho5 (Lechner et al., 2011). 2333/4254 (54.8%) genes in the B2 genome were unique and did not have orthology to sequences in other reference Atribacteria genomes.

Atribacteria community structure partitioning: Community structure for Atribacteria OTUs from all depths was analyzed in the following ways:

1) *Tetra-nucleotide compositions*: Normalized tetranucleotide composition was calculated for each Atribacteria OTU (non-rarified) from all depths and was used for a principal component analysis (PCA) in R.

2) *Rank abundance correlations*: OTU abundances based on the rarified Atribacteria OTU table (previously generated during diversity analysis) were $\log_{10}$ transformed in R. Pairwise

Pearson correlations were calculated between depths using the rcorr function and plotted using the chart.Correlation function from the PerformanceAnalytics package in R. A heatmap with hierarchical clustering of depths was generated using these correlations with heatmap function in R.

3) *NMDS analysis*: OTU abundance from the rarified Atribacteria OTU table (previously generated during diversity analysis) was used for NMDS analysis after square root transformation and calculation of the Bray-Curtis dissimilarity metric, all via the metaMDS function using the Vegan package in R. After examination of the shepard plot for scatter around the regression line, the NMDS plot was created showing individual OTUs and the midpoint for whole communities. A hierarchical clustering dendrogram was generated using the Bray-Curtis dissimilarities.

3.3 Results and Discussion

Atribacteria community abundance in the methane hydrate sediment. Targeted 16S rRNA gene sequencing demonstrated that the relative abundance of Atribacteria increased with depth, from 15% in the near surface, to 86% in the deeper depths in the methane hydrate sediments (Figure 10). This trend is consistent with other studies suggesting higher Atribacteria abundance with increase in methane concentration in the sediments (Carr et al., 2015).

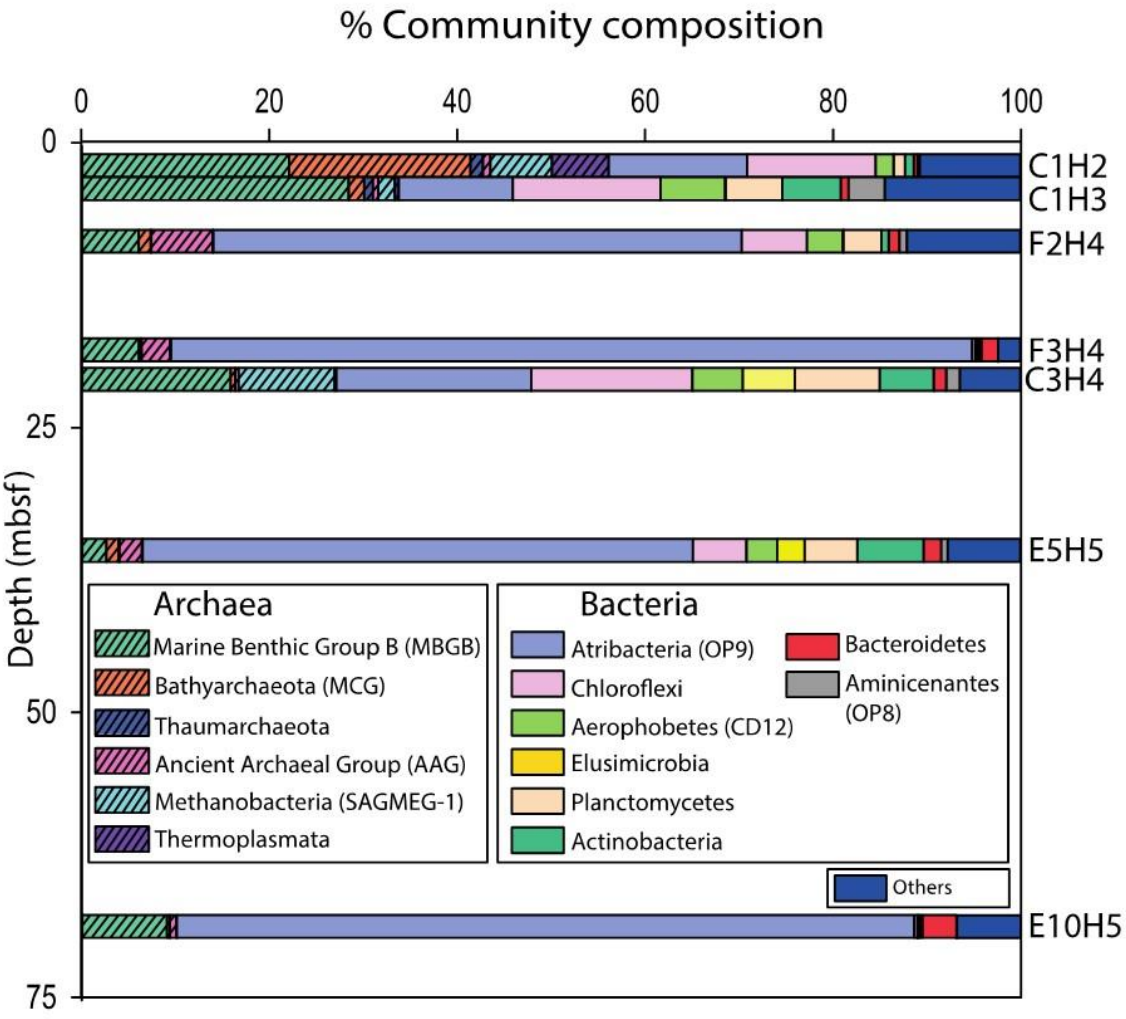


Figure 10 – Microbial abundance estimates in sediment samples from ODP site 1244.

Abundances are estimated through 16S amplicons covering depths from 2 to 68.55 m below seafloor (mbsf).

Atribacteria MAG from the sediment metagenomes. We assembled community metagenomes from seven sediment depths from ODP site 1244 and annotated the contig sequences for coding and non-coding genomic elements. The overall assembly size for all depths varied between 39.8 – 195.4 Mbp and demonstrated high assembly/read recruitment rates varying between 75-93.42% (Table 4). Later using metagenomic binning of the assembled contig sequences, we extracted several genomes (MAGs) from all depths. Even

after discarding genome candidates with higher than 10% contamination, these MAGs demonstrated highly variable levels of completeness when assessed by the presence of essential single copy marker genes. A list of MAGs showing the best genome candidates based on lower contamination and higher completeness are shown in Table 5. From these sediment metagenomes, we demonstrate extraction of two Atribacteria MAGs, one from an intermediate (F2H4, 8.6 mbsf) depth labelled F2H4-B2 and other from the deepest (E10H5, 68.55 mbsf) depth labelled E10H5-B2. In addition to Atribacteria MAGs, we found several other interesting genomes that are still under-characterized in the scientific literature, for example genomes affiliated with the Lokiarchaea and Hadesarchaea. This opens a genomic window to a much better understanding of the functional potential of such understudied microbial lineages that constitute the so called 'microbial dark matter' (Rinke et al., 2013; Marcy et al., 2007).

Table 4 – Assembly statistics for ODP site 1244 sediment metagenomes.

Dataset	Depth (m)	QC Reads ($x10^6$)	Contigs ($x10^3$)	N50 (bp)	L50 (#)	Total length (Mbp)	Largest contig (Kbp)	Unique genes ($x10^3$)	Total genes ($x10^3$)	Reads aligned (%)
C1H2	2.00	21.8	35.8	2,605	6,125	66.7	73.5	91.4	91.9	93.42
C1H3	3.55	23.5	126.7	2,041	24,423	195.4	47.3	277.4	278.0	75.36
F2H4	8.60	24.2	43.7	2,330	7,511	72.8	103.7	101.0	101.2	91.17
F3H4	18.10	19.2	25.3	3,953	3,324	57.1	232.4	73.1	73.8	93.60
C3H4	20.69	19.9	71.8	2,638	11,911	129.4	349.3	174.7	175.2	84.51
E5H5	35.65	20.4	82.5	2,209	15,482	133.9	43.1	186.6	187.0	83.49
E10H5	68.55	18.2	21.1	2,777	3,865	39.8	41.0	53.5	53.7	93.03

Table 5 – Genome statistics for MAGs extracted form ODP 1244 sediment metagenomes in all depths.

Genome statistics for MAGs with less than 10% contamination and highest completeness estimates are shown. Atribacteria MAGs extracted are labelled F2H4-B2 and E10H5-B2.

MAG ID	Completeness	Contamination	Strain Heterogenity	LCA Taxonomy	Contigs	Genome Size (Kbp)	GC (%)	CDS	rRNA	16S	rpoB	rpoB Top Hits
C1H2-B8	48.07	4.01	0	Firmicutes	135	591.5	45.41 ± 1.45	594			1	Moorella glycerini(57.62)
C1H2-B1	41.44	1.72	0	Firmicutes	146	773.8	42.62 ± 1.43	738				
C1H3-B19	60.6	1.72	0	Bacteria	226	1724.3	52.79 ± 1.71	1686				
C1H3-B24	51.8	6.9	0	Desulfobacteraceae; unc	724	2128.6	46.24 ± 2.01	2001				
C1H3-B37	49.92	0	0	Bacteroidetes	419	1469.3	47.00 ± 2.33	1426				
C1H3-B2	47.69	10	18.75	Clostridia	269	1210.7	33.97 ± 2.60	1078			1	Thermoanaerobacterales bacterium 50_218(69.07)
C1H3-B32	42.79	1.03	50	unc Parcubacteria	72	333.2	37.24 ± 1.48	350			1	Parcubacteria group bacterium GW2011_GWC1_38_6(93.39)
C1H3-B47	42.62	2.22	0	Firmicutes	139	495.0	44.58 ± 2.09	503				
F2H4-B2	79.07	4.44	50	Clostridia	406	2136.0	40.48 ± 1.61	2041	6	1	2	Atribacteria bacterium 34_128(42.57); Moorella sp. 60_41(63.54)
F2H4-B14	69.76	8.05	0	Acidobacteria	561	1980.9	41.00 ± 2.19	1908	2		1	Elstera litoralis(79.82)
F2H4-B12	45.8	7.18	0	Bacteroidales	308	1318.2	37.64 ± 1.59	1263			2	candidate division WOR_3 bacterium SM23_42(92.44,82.8)
F3H4-B3	89.83	5.1	33.33	Clostridia	132	1988.6	40.92 ± 1.78	1758	3	1	1	Elusimicrobium minutum(57.18)
F3H4-B2	81.03	0	0	Clostridia	47	2140.0	34.30 ± 1.54	2077	4	1		
F3H4-B6	60.1	1.95	0	Methanosarcina	165	1093.4	42.82 ± 1.63	1299			1	Methanolobus sp. T82-4(75.94)
C3H4-B1	88.1	4.01	16.67	Euryarchaeota; can div MSBL1	20	1294.5	50.40 ± 1.43	1460	3	1	1	Hadesarchaea archaeon YNP_45(81.91)
C3H4-B19	77.26	6.22	14.29	Bacteroides	750	3577.9	37.77 ± 2.10	2889	1		1	Bacteroides(82.67)
C3H4-B7	64.96	6.92	21.05	Euryarchaeota; can div MSBL1	141	1071.5	53.03 ± 2.88	1243	1		2	Hadesarchaea archaeon(82.41,82.82)
C3H4-B23	59.01	0.71	0	Bacteria	349	1438.0	41.40 ± 1.89	1342			1	Elstera litoralis(79.82,partial)
C3H4-B5	56.9	1.67	0	Clostridia	85	998.1	41.99 ± 1.12	881	3	1		
C3H4-B25	43.69	2.17	100	Bacteria	240	1062.2	52.51 ± 1.93	1052	1			
C3H4-B11	42.14	0.49	0	Firmicutes	213	828.2	41.36 ± 1.59	826				
E5H5-B16	52.68	3.45	0	Firmicutes	497	1561.6	41.20 ± 2.07	1514				
E5H5-B17	49.45	1.28	0	Thaumarchaeota	249	1307.2	58.38 ± 2.52	1467			1	miscellaneous Crenarchaeota group-15 archaeon DG-45(83.94)
E5H5-B30	41.44	2.4	28.57	Bradyrhizobium	431	3284.5	62.93 ± 1.39	3385	1	1		
E5H5-B19	40.5	0	0	Spirochaeta	390	1522.4	47.03 ± 2.59	1452				
E10H5-B2	68.88	1.85	100	Clostridia + Proteobacteria	912	4055.3	34.72 ± 2.21	4254			1	Atribacteria bacterium 34_128(94.47)
E10H5-B12	47.46	8.55	0	Clostridia + Archaea	1343	4158.4	31.04 ± 2.50	4283			3	Lokiarchaeum sp. GC14_75(94.92,92.44,91.89)

Atribacteria F2H4-B2 MAG was 79% complete with 4.5% contamination according to the marker gene estimates. However, only 50% strain heterogeneity suggests that the other half of the duplicate marker genes are on contigs that could be a part of another genome. This contamination is apparent through the presence of a second copy of marker gene *rpoB* that matches to a *Moorella* genome. While the F2H4-B2 MAG demonstrated foreign contamination, Atribacteria MAG E10H5-B2 demonstrated no such contamination in its genome. E10H5-B2 MAG reported 70% completeness and a lower

overall contamination rate of 1.85% by maker gene estimates. That, combined with the observed 100% strain heterogeneity, suggests that the duplicity in the marker genes is due to a strain level differentiation. The genome recovered for E10H5-B2 demonstrated an overall size of about 4 Mb at 70% completeness, suggesting an expected genome size of about 5.8 Mb. This estimate is significantly higher that other Atribacteria genome sizes (Table 3). This increase in genome size along with a high strain heterogeneity leads us to believe that Atribacteria has high diversity in this community and that this MAG is an ensemble of the Atribacteria population rather than a single genome. Apart from these characteristics, E10H5-B2 demonstrates similarly low GC content (35 ± 2%, Table 5) compared to other JS1 type Atribacteria (35 – 38%, Table 3) but lacks a recovered 16S rRNA gene. Its *rpoB* gene shows 94% similarity to Atribacteria bacterium 34_128 from an oil reservoir. Phylogenetic reconstruction based on conserved protein-coding markers in the E10H5-B2 MAG, known Atribacteria SAGs/MAGs and several other affiliated bacterial phyla with cultivated representatives, show strong support for designation of Atribacteria as a phylum level lineage in the Bacteria (Figure 11 A). E10H5-B2 MAG clusters with other SAGs/MAGs belonging to the JS1-1 genus within the JS1 lineage, with the closest relative being the Atribacteria bacterium 34_128 from an oil reservoir environment (Figure 11 B).

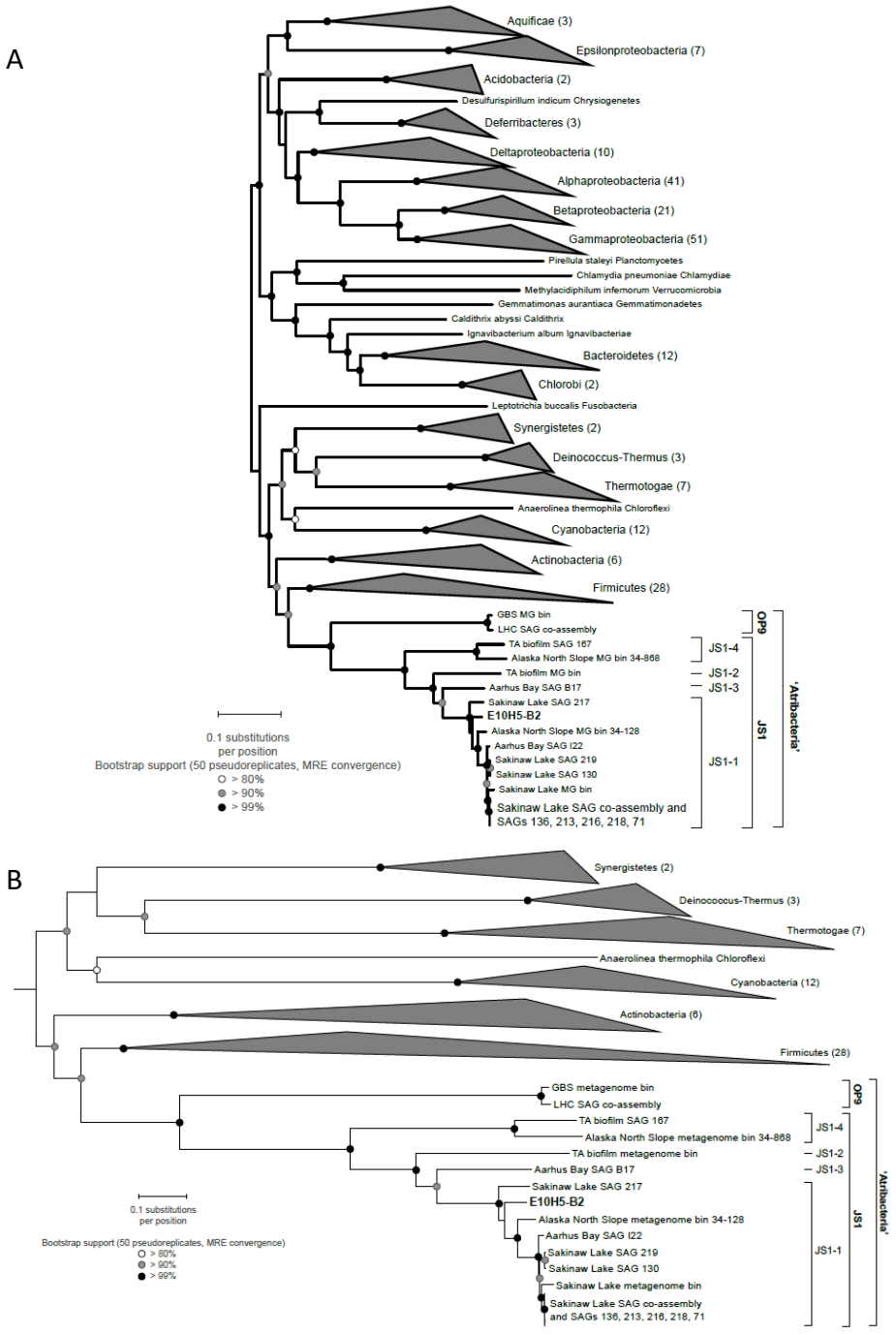


Figure 11 – Phylogenetic reconstruction for MAG E10H5-B2.

A) Phylogenetic analysis of Atribacteria MAGs and SAGs and other Bacteria. B) Pruned view showing E10H5-B2 MAG included in Atribacteria JS1-1 clade. Maximum likelihood phylogeny inferred with RAxML (Stamatakis, 2006) using a concatenated alignment of 6-69 single copy conserved marker genes. The number of organisms represented by each wedge is indicated in parentheses.

Atribacteria community diversity in the methane hydrate sediments. 16S rRNA amplicon sequences were used to estimate the microdiversity in the Atribacteria population. Candidate phylum Atribacteria is divided into three classes. OP9, which covers classes 2 and 3, including Atribacteria from hot geothermal environments (Nobu et al., 2016; Carr et al., 2015; Yarza et al., 2014). Other OP9 candidates used in this study are from thermal bioreactors. The JS-1 clade is divided into three families with most candidates coming from different marine environments (Nobu et al., 2016; Carr et al., 2015; Yarza et al., 2014). JS1 Family 3 is known by a single candidate that comes from a marine hydrothermal vent environment. JS1 Family 2 has five genera, of which one genus is identified from oil reservoir fluid and the other four from terrestrial sediments. JS1 Family 1 is again broken further into 5 genera. Candidates in genus 5, 3 and 2 are from multiple sources (oil reservoirs, hot springs and terrestrial sediments), while candidates in genus 1 and 4 are predominantly from marine environments (Yarza et al., 2014). Genus 1 displays the maximum representation when it comes to number of sequences available, most of them from either methane hydrate seeps or mud volcanoes (Figure 12). 16S rRNA gene sequences clustered and identified as Atribacteria from the E10H5 (68.55mbsf) sample, nearly all, were placed phylogenetically along branches in genus 1, along with sequences coming from a hydrate seep environment (Figure 12). This confirms that the Atribacteria community in E10H5 sample is from the JS1-1 clade and is consistent with the finding that E10H5-B2 MAG clusters within the JS1-1 genus.

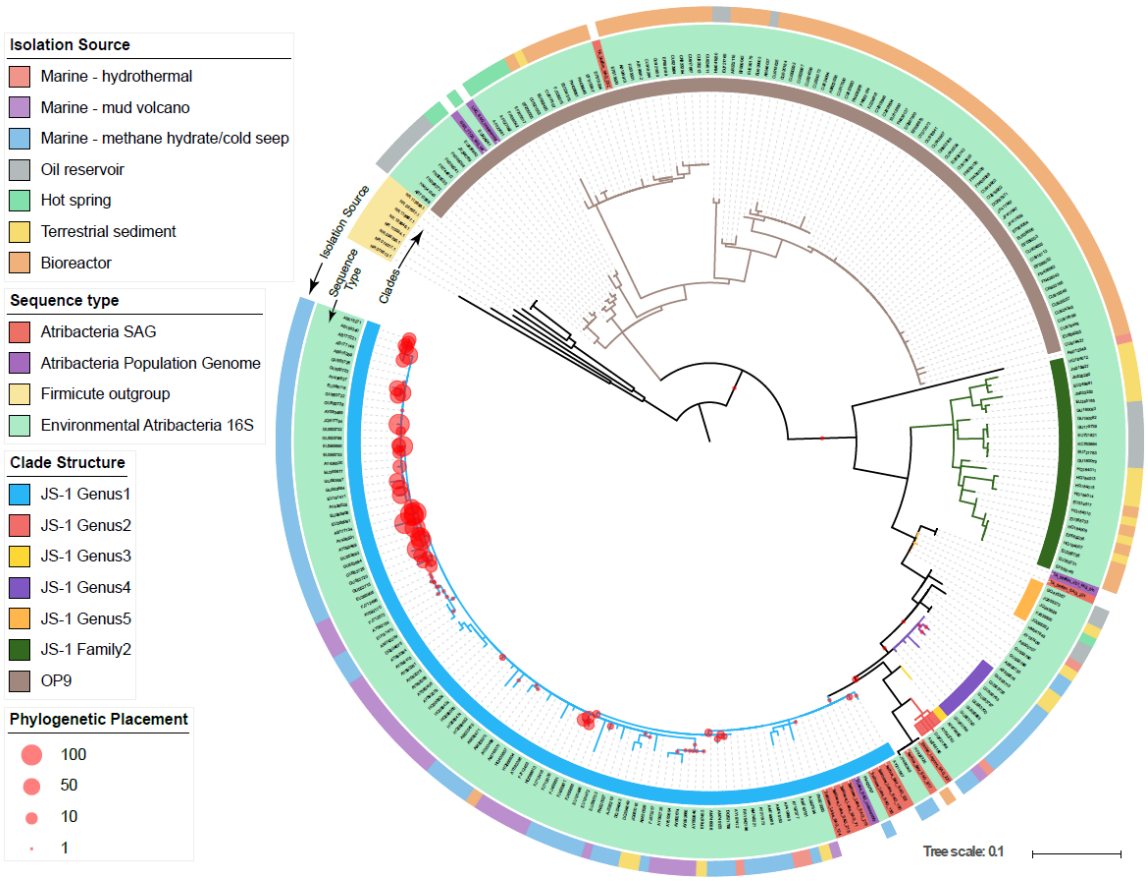


Figure 12 – Phylogenetic placement of Atribacteria 16S amplicons on known Atribacteria clades.

Maximum likelihood phylogeny inferred with RAxML (Stamatakis, 2006) using full-length 16S rRNA sequences from known Atribacteria clades. Maximum probabilistic (up to a max of 7 placements per OTU) phylogenetic placements (red circles) for 230 Atribacteria OTU from E10H5 68.55 mbsf sample was performed using EPA in RAxML.

Atribacteria OTUs recovered from the E10H5 sample, when represented in a phylogenetic reconstruction among other Atribacteria 16S sequences, demonstrate rich microdiversity within the JS1-1 clade (Figure 13). This demonstration of microdiversity also spanned a wide range of relatedness within the JS1 Genus 1. On average, the OTUs recovered from E10H5 share 92% 16S rRNA gene similarity, which is roughly equivalent to separation at the genus/species level. In some cases, however, similarity between OTUs dropped to 89%. There was wide variation in the relative abundance of the recovered

OTUs, with one (OTU-007) comprising 68.5% of total Atribacteria sequences from E10H5 (Figure 13). The composition of the Atribacteria community varied with depth, and the community from the shallowest depth (2 mbsf) was notably distinct from that of the deeper sediment (Figure 14).

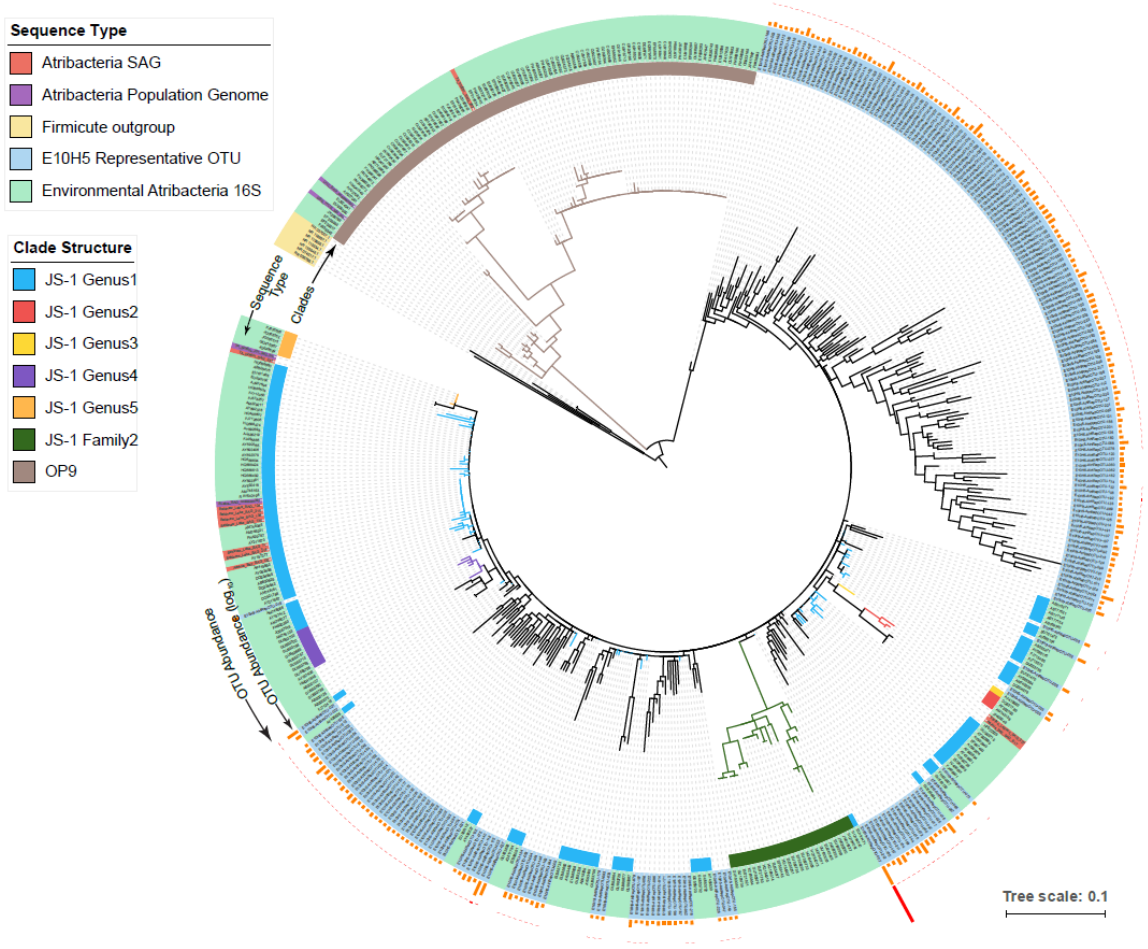


Figure 13 – Phylogenetic reconstruction of Atribacteria 16S amplicons with known Atribacteria clades.

Maximum likelihood phylogeny inferred with RAxML (Stamatakis, 2006) using V3-V4 region of 16S rRNA sequences from 230 Atribacteria OTUs recovered from E10H5 68.55 mbsf sample along with known Atribacteria sequences.

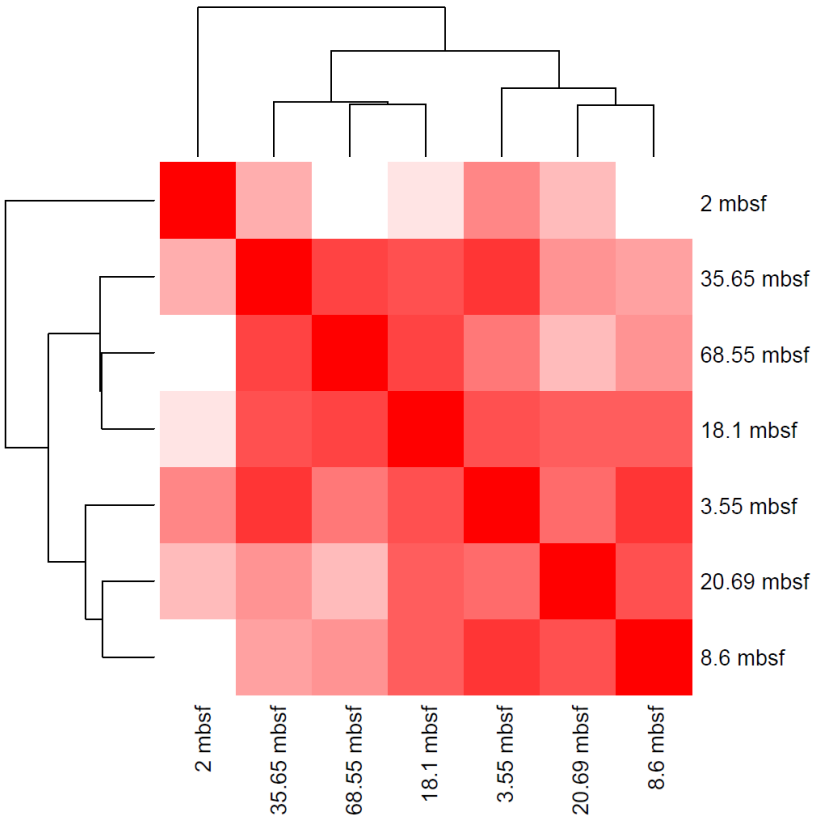


Figure 14 – Clustering of Atribacteria community by their abundance.

Heat map with colors representing pairwise Pearson correlation and a hierarchical clustering dendrogram showing sample (depth) groupings. OTU abundances were log transformed and OTUs with 0 abundance (absence) are ignored.

3.4 Conclusions

Atribacteria is a globally distributed candidate phylum found mostly in organic rich anaerobic environments. The lineage sub-division of Atribacteria in OP9 and JS1 differentiates clearly by their source environment, where OP9 are mostly found in geothermal springs while the JS1 are abundant in marine methane hydrate cold sediments and oil reservoirs (Nobu et al., 2016). While the OP9 clade is well characterized for its genome potential and diversity, the JS1 clade needs more investigation given its diversified clade structure. Here, we take advantage of the complementarity of metagenomics and 16S

rRNA gene surveys to assemble a nearly complete genome of JS1 type candidate bacterial phylum Atribacteria along with demonstration of the Atribacteria population diversity in the microbial communities from a methane hydrate-bearing marine sediment. 16S rRNA gene phylogenetics firmly places Atribacteria community members from the E10H5 sample within JS1-1 genus of this candidate phylum whereas whole genome phylogenetics firmly places the E10H5-B2 MAG under the same genus. The recovered JS1 Atribacteria genome opens the opportunity to better understand niche-specific adaptations in Atribacteria from such extreme environments. A complete understanding of such adaptations would illustrate the reasons behind high abundances of Atribacteria in such communities.

REFERENCES

Alldredge, A. L., & Cohen, Y. (1987). Can microscale chemical patches persist in the sea? Microelectrode study of marine snow, fecal pellets. Science, 235(4789), 689-691.

Alldredge, A. L., & Silver, M. W. (1988). Characteristics, dynamics and significance of marine snow. Progress in oceanography, 20(1), 41-82.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. Journal of molecular biology, 215(3), 403-410.

Anderson, J. J., & Devol, A. H. (1973). Deep water renewal in Saanich Inlet, an intermittently anoxic basin. Estuarine and Coastal Marine Science, 1(1), 1-10.

Anderson, P. N., Hume, M. E., Byrd, J. A., Hernandez, C., Stevens, S. M., Stringfellow, K., & Caldwell, D. J. (2010). Evaluation of repetitive extragenic palindromic-polymerase chain reaction and denatured gradient gel electrophoresis in identifying Salmonella serotypes isolated from processed turkeys. Poultry science, 89(6), 1293-1300.

Andersson, A. F., & Banfield, J. F. (2008). Virus population dynamics and acquired virus resistance in natural microbial communities. Science, 320(5879), 1047-1050.

Arber, W. (1979). Promotion and limitation of genetic exchange. Experientia, 35(3), 287-293.

Atmos, J. (2009, September 15). Diagram of the possible mechanism for CRISPR. Retrieved from https://en.wikipedia.org/wiki/File:Crispr.png

Baltimore, D., Berg, P., Botchan, M., Carroll, D., Charo, R. A., Church, G., ... & Greely, H. T. (2015). A prudent path forward for genomic engineering and germline gene modification. Science, 348(6230), 36-38.

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A. A., Dvorkin, M., Kulikov, A. S., ... & Pyshkin, A. V. (2012). SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. Journal of computational biology, 19(5), 455-477.

Barrangou, R. (2013). CRISPR-Cas systems and RNA-guided interference. Wiley Interdisciplinary Reviews: RNA, 4(3), 267-278.

Barrangou, R., Fremaux, C., Deveau, H., Richards, M., Boyaval, P., Moineau, S., ... & Horvath, P. (2007). CRISPR provides acquired resistance against viruses in prokaryotes. Science, 315(5819), 1709-1712.

Barrangou, R., & Horvath, P. (2012). CRISPR: new horizons in phage resistance and strain identification. Annual review of food science and technology, 3, 143-162.

Barrangou, R., & Marraffini, L. A. (2014). CRISPR-Cas systems: prokaryotes upgrade to adaptive immunity. Molecular cell, 54(2), 234-244.

Beloglazova, N., Brown, G., Zimmerman, M. D., Proudfoot, M., Makarova, K. S., Kudritska, M., ... & Koonin, E. V. (2008). A novel family of sequence-specific endoribonucleases associated with the clustered regularly interspaced short palindromic repeats. Journal of Biological Chemistry, 283(29), 20361-20371.

Bikard, D., Euler, C. W., Jiang, W., Nussenzweig, P. M., Goldberg, G. W., Duportet, X., ... & Marraffini, L. A. (2014). Exploiting CRISPR-Cas nucleases to produce sequence-specific antimicrobials. Nature biotechnology, 32(11), 1146.

Bland, C., Ramsey, T. L., Sabree, F., Lowe, M., Brown, K., Kyrpides, N. C., & Hugenholtz, P. (2007). CRISPR recognition tool (CRT): a tool for automatic detection of clustered regularly interspaced palindromic repeats. BMC bioinformatics, 8(1), 209.

Boch, J., Scholze, H., Schornack, S., Landgraf, A., Hahn, S., Kay, S., ... & Bonas, U. (2009). Breaking the code of DNA binding specificity of TAL-type III effectors. Science, 326(5959), 1509-1512.

Bolderson, E., Richard, D. J., Zhou, B. B. S., & Khanna, K. K. (2009). Recent advances in cancer therapy targeting proteins involved in DNA double-strand break repair. Clinical Cancer Research, 15(20), 6314-6320.

Bolotin, A., Quinquis, B., Sorokin, A., & Ehrlich, S. D. (2005). Clustered regularly interspaced short palindrome repeats (CRISPRs) have spacers of extrachromosomal origin. Microbiology, 151(8), 2551-2561.

Brouns, S. J., Jore, M. M., Lundgren, M., Westra, E. R., Slijkhuis, R. J., Snijders, A. P., ... & Van Der Oost, J. (2008). Small CRISPR RNAs guide antiviral defense in prokaryotes. Science, 321(5891), 960-964.

Bryant, J. A., Stewart, F. J., Eppley, J. M., & DeLong, E. F. (2012). Microbial community phylogenetic and trait diversity declines with depth in a marine oxygen minimum zone. Ecology, 93(7), 1659-1673.

Canfield, D. E., Stewart, F. J., Thamdrup, B., De Brabandere, L., Dalsgaard, T., Delong, E. F., ... & Ulloa, O. (2010). A cryptic sulfur cycle in oxygen-minimum–zone waters off the Chilean coast. Science, 330(6009), 1375-1378.

Caporaso, J. G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F. D., Costello, E. K., ... & Huttley, G. A. (2010). QIIME allows analysis of high-throughput community sequencing data. Nature methods, 7(5), 335.

Carr, S. A., Orcutt, B. N., Mandernack, K. W., & Spear, J. R. (2015). Abundant Atribacteria in deep marine sediment from the Adélie Basin, Antarctica. Frontiers in microbiology, 6, 872.

Carroll, D. (2011). Genome engineering with zinc-finger nucleases. Genetics, 188(4), 773-782.

Cassman, N., Prieto-Davó, A., Walsh, K., Silva, G. G., Angly, F., Akhter, S., ... & Willner, D. (2012). Oxygen minimum zones harbour novel viral communities with low diversity. Environmental microbiology, 14(11), 3043-3065.

Chibani-Chennoufi, S., Bruttin, A., Dillmann, M. L., & Brüssow, H. (2004). Phage-host interaction: an ecological perspective. Journal of bacteriology, 186(12), 3677-3686.

Cho, B. C., & Azam, F. (1988). Major role of bacteria in biogeochemical fluxes in the ocean's interior. Nature, 332(6163), 441.

Cho, S. W., Kim, S., Kim, Y., Kweon, J., Kim, H. S., Bae, S., & Kim, J. S. (2014). Analysis of off-target effects of CRISPR/Cas-derived RNA-guided endonucleases and nickases. Genome research, 24(1), 132-141.

Chouari, R., Le Paslier, D., Daegelen, P., Ginestet, P., Weissenbach, J., & Sghir, A. (2005). Novel predominant archaeal and bacterial groups revealed by molecular analysis of an anaerobic sludge digester. Environmental Microbiology, 7(8), 1104-1115.

Citorik, R. J., Mimee, M., & Lu, T. K. (2014). Sequence-specific antimicrobials using efficiently delivered RNA-guided nucleases. Nature biotechnology, 32(11), 1141.

Clasen, J. L., Brigden, S. M., Payet, J. P., & Suttle, C. A. (2008). Evidence that viral abundance across oceans and lakes is driven by different biological factors. Freshwater Biology, 53(6), 1090-1100.

Conesa, A., Götz, S., García-Gómez, J. M., Terol, J., Talón, M., & Robles, M. (2005). Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics, 21(18), 3674-3676.

Cong, L., Ran, F. A., Cox, D., Lin, S., Barretto, R., Habib, N., ... & Zhang, F. (2013). Multiplex genome engineering using CRISPR/Cas systems. Science, 1231143.

Conley, D. J., Humborg, C., Rahm, L., Savchuk, O. P., & Wulff, F. (2002). Hypoxia in the Baltic Sea and basin-scale changes in phosphorus biogeochemistry. Environmental science & technology, 36(24), 5315-5320.

Cornu, T. I., & Cathomen, T. (2010). Quantification of zinc finger nuclease-associated toxicity. In Engineered Zinc Finger Proteins (pp. 237-245). Humana Press, Totowa, NJ.

Costa, K. C., Navarro, J. B., Shock, E. L., Zhang, C. L., Soukup, D., & Hedlund, B. P. (2009). Microbiology and geochemistry of great boiling and mud hot springs in the United States Great Basin. Extremophiles, 13(3), 447-459.

Cradick, T. J., Fine, E. J., Antico, C. J., & Bao, G. (2013). CRISPR/Cas9 systems targeting β-globin and CCR5 genes have substantial off-target activity. Nucleic acids research, 41(20), 9584-9592.

De Corte, D., Sintes, E., Yokokawa, T., Reinthaler, T., & Herndl, G. J. (2012). Links between viruses and prokaryotes throughout the water column along a North Atlantic latitudinal transect. The ISME journal, 6(8), 1566.

DeLong, E. F., Franks, D. G., & Alldredge, A. L. (1993). Phylogenetic diversity of aggregate-attached vs. free-living marine bacterial assemblages. Limnology and Oceanography, 38(5), 924-934.

DeLong, E. F., Preston, C. M., Mincer, T., Rich, V., Hallam, S. J., Frigaard, N. U., ... & Chisholm, S. W. (2006). Community genomics among stratified microbial assemblages in the ocean's interior. Science, 311(5760), 496-503.

Deltcheva, E., Chylinski, K., Sharma, C. M., Gonzales, K., Chao, Y., Pirzada, Z. A., ... & Charpentier, E. (2011). CRISPR RNA maturation by trans-encoded small RNA and host factor RNase III. Nature, 471(7340), 602.

Deveau, H., Barrangou, R., Garneau, J. E., Labonté, J., Fremaux, C., Boyaval, P., ... & Moineau, S. (2008). Phage response to CRISPR-encoded resistance in Streptococcus thermophilus. Journal of bacteriology, 190(4), 1390-1400.

Deveau, H., Garneau, J. E., & Moineau, S. (2010). CRISPR/Cas system and its role in phage-bacteria interactions. Annual review of microbiology, 64, 475-493.

Dodsworth, J. A., Blainey, P. C., Murugapiran, S. K., Swingley, W. D., Ross, C. A., Tringe, S. G., ... & Quake, S. R. (2013). Single-cell and metagenomic analyses indicate a fermentative and saccharolytic lifestyle for members of the OP9 lineage. Nature communications, 4, 1854.

Doudna, J. A., & Charpentier, E. (2014). The new frontier of genome engineering with CRISPR-Cas9. Science, 346(6213), 1258096.

Eloe, E. A., Shulse, C. N., Fadrosh, D. W., Williamson, S. J., Allen, E. E., & Bartlett, D. H. (2011). Compositional differences in particle-associated and free-living microbial assemblages from an extreme deep-ocean environment. Environmental microbiology reports, 3(4), 449-458.

Elshahed, M. S., Youssef, N. H., Spain, A. M., Sheik, C., Najar, F. Z., Sukharnikov, L. O., ... & Krumholz, L. R. (2008). Novelty and uniqueness patterns of rare members of the soil biosphere. Applied and Environmental Microbiology, 74(17), 5422-5428.

Faith, D. P. (1992). Conservation evaluation and phylogenetic diversity. Biological conservation, 61(1), 1-10.

Fineran, P. C., & Charpentier, E. (2012). Memory of viral infections by CRISPR-Cas adaptive immune systems: acquisition of new information. Virology, 434(2), 202-209.

Fu, Y., Foden, J. A., Khayter, C., Maeder, M. L., Reyon, D., Joung, J. K., & Sander, J. D. (2013). High-frequency off-target mutagenesis induced by CRISPR-Cas nucleases in human cells. Nature biotechnology, 31(9), 822.

Fuhrman, J. A., & Suttle, C. A. (1993). Viruses in marine planktonic systems. Oceanography, 6(2), 51-63.

Gabriel, R., Lombardo, A., Arens, A., Miller, J. C., Genovese, P., Kaeppel, C., ... & Holmes, M. C. (2011). An unbiased genome-wide analysis of zinc-finger nuclease specificity. Nature biotechnology, 29(9), 816.

Garneau, J. E., Dupuis, M. È., Villion, M., Romero, D. A., Barrangou, R., Boyaval, P., ... & Moineau, S. (2010). The CRISPR/Cas bacterial immune system cleaves bacteriophage and plasmid DNA. Nature, 468(7320), 67.

Gasiunas, G., Barrangou, R., Horvath, P., & Siksnys, V. (2012). Cas9–crRNA ribonucleoprotein complex mediates specific DNA cleavage for adaptive immunity in bacteria. Proceedings of the National Academy of Sciences, 109(39), E2579-E2586.

Gittel, A., Sørensen, K. B., Skovhus, T. L., Ingvorsen, K., & Schramm, A. (2009). Prokaryotic community structure and sulfate reducer activity in water from high-temperature oil reservoirs with and without nitrate treatment. Applied and Environmental Microbiology, 75(22), 7086-7096.

Grossart, H. P., Kiørboe, T., Tang, K. W., Allgaier, M., Yam, E. M., & Ploug, H. (2006). Interactions between marine snow and heterotrophic bacteria: aggregate formation and microbial dynamics. Aquatic microbial ecology, 42, 19-26.

Gurevich, A., Saveliev, V., Vyahhi, N., & Tesler, G. (2013). QUAST: quality assessment tool for genome assemblies. Bioinformatics, 29(8), 1072-1075.

Hale, C. R., Majumdar, S., Elmore, J., Pfister, N., Compton, M., Olson, S., ... & Terns, M. P. (2012). Essential features and rational design of CRISPR RNAs that function with the Cas RAMP module complex to cleave RNAs. Molecular cell, 45(3), 292-302.

Hale, C. R., Zhao, P., Olson, S., Duff, M. O., Graveley, B. R., Wells, L., ... & Terns, M. P. (2009). RNA-guided RNA cleavage by a CRISPR RNA-Cas protein complex. Cell, 139(5), 945-956.

Hara S., Koike I., Terauchi K., Kamiya H., Tanoue E. (1996) Abundance of viruses in deep oceanic waters. Marine Ecology Progress Series, 145, 269-277.

Hara, S., Terauchi, K., & Koike, I. (1991). Abundance of viruses in marine waters: assessment by epifluorescence and transmission electron microscopy. Applied and Environmental Microbiology, 57(9), 2731-2734.

Harris, J. K., Caporaso, J. G., Walker, J. J., Spear, J. R., Gold, N. J., Robertson, C. E., ... & Marshall, P. (2013). Phylogenetic stratigraphy in the Guerrero Negro hypersaline microbial mat. The ISME journal, 7(1), 50.

Hockemeyer, D., Wang, H., Kiani, S., Lai, C. S., Gao, Q., Cassady, J. P., ... & Zeitler, B. (2011). Genetic engineering of human pluripotent cells using TALE nucleases. Nature biotechnology, 29(8), 731.

Hollibaugh, J. T., Wong, P. S., & Murrell, M. C. (2000). Similarity of particle-associated and free-living bacterial communities in northern San Francisco Bay, California. Aquatic Microbial Ecology, 21(2), 103-114.

Horvath, P., & Barrangou, R. (2010). CRISPR/Cas, the immune system of bacteria and archaea. Science, 327(5962), 167-170.

Hsu, P. D., Lander, E. S., & Zhang, F. (2014). Development and applications of CRISPR-Cas9 for genome engineering. Cell, 157(6), 1262-1278.

Hsu, P. D., Scott, D. A., Weinstein, J. A., Ran, F. A., Konermann, S., Agarwala, V., ... & Cradick, T. J. (2013). DNA targeting specificity of RNA-guided Cas9 nucleases. Nature biotechnology, 31(9), 827.

Hu, W., Kaminski, R., Yang, F., Zhang, Y., Cosentino, L., Li, F., ... & Mo, X. (2014). RNA-directed gene editing specifically eradicates latent and prevents new HIV-1 infection. Proceedings of the National Academy of Sciences, 111(31), 11461-11466.

Hugenholtz, P., Pitulle, C., Hershberger, K. L., & Pace, N. R. (1998). Novel division level bacterial diversity in a Yellowstone hot spring. Journal of bacteriology, 180(2), 366-376.

Hunt, D. E., David, L. A., Gevers, D., Preheim, S. P., Alm, E. J., & Polz, M. F. (2008). Resource partitioning and sympatric differentiation among closely related bacterioplankton. Science, 320(5879), 1081-1085.

Huson, D. H., Auch, A. F., Qi, J., & Schuster, S. C. (2007). MEGAN analysis of metagenomic data. Genome research, 17(3), 377-386.

Hyatt, D., Chen, G. L., LoCascio, P. F., Land, M. L., Larimer, F. W., & Hauser, L. J. (2010). Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC bioinformatics, 11(1), 119.

Inagaki, F., Nunoura, T., Nakagawa, S., Teske, A., Lever, M., Lauer, A., ... & Nealson, K. H. (2006). Biogeographical distribution and diversity of microbes in methane hydrate-bearing deep marine sediments on the Pacific Ocean Margin. Proceedings of the National Academy of Sciences of the United States of America, 103(8), 2815-2820.

Inagaki, F., Suzuki, M., Takai, K., Oida, H., Sakamoto, T., Aoki, K., ... & Horikoshi, K. (2003). Microbial communities associated with geological horizons in coastal subseafloor sediments from the Sea of Okhotsk. Applied and Environmental Microbiology, 69(12), 7224-7235.

Ishino, Y., Shinagawa, H., Makino, K., Amemura, M., & Nakata, A. (1987). Nucleotide sequence of the iap gene, responsible for alkaline phosphatase isozyme conversion in Escherichia coli, and identification of the gene product. Journal of bacteriology, 169(12), 5429-5433.

Jansen, R., Embden, J., Gaastra, W., & Schouls, L. (2002). Identification of genes that are associated with DNA repeats in prokaryotes. Molecular microbiology, 43(6), 1565-1575.

Jiang, W., Bikard, D., Cox, D., Zhang, F., & Marraffini, L. A. (2013). RNA-guided editing of bacterial genomes using CRISPR-Cas systems. Nature biotechnology, 31(3), 233.

Jinek, M., Chylinski, K., Fonfara, I., Hauer, M., Doudna, J. A., & Charpentier, E. (2012). A programmable dual-RNA–guided DNA endonuclease in adaptive bacterial immunity. science, 1225829.

Jinek, M., East, A., Cheng, A., Lin, S., Ma, E., & Doudna, J. (2013). RNA-programmed genome editing in human cells. elife, 2.

Jones, P., Binns, D., Chang, H. Y., Fraser, M., Li, W., McAnulla, C., ... & Pesseat, S. (2014). InterProScan 5: genome-scale protein function classification. Bioinformatics, 30(9), 1236-1240.

Jørgensen, B. B. (1982). Ecology of the bacteria of the sulphur cycle with special reference to anoxic—oxic interface environments. Phil. Trans. R. Soc. Lond. B, 298(1093), 543-561.

Kanehisa, M., Sato, Y., & Morishima, K. (2016). BlastKOALA and GhostKOALA: KEGG tools for functional characterization of genome and metagenome sequences. Journal of molecular biology, 428(4), 726-731.

Kang, D. D., Froula, J., Egan, R., & Wang, Z. (2015). MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities. PeerJ, 3, e1165.

Karstensen, J., Stramma, L., & Visbeck, M. (2008). Oxygen minimum zones in the eastern tropical Atlantic and Pacific oceans. Progress in Oceanography, 77(4), 331-350.

Katoh, K., & Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Molecular biology and evolution, 30(4), 772-780.

Kellogg, C. T., & Deming, J. W. (2009). Comparison of free-living, suspended particle, and aggregate-associated bacterial and archaeal communities in the Laptev Sea. Aquatic Microbial Ecology, 57(1), 1-18.

Kobayashi, H., Endo, K., Sakata, S., Mayumi, D., Kawaguchi, H., Ikarashi, M., ... & Sato, K. (2012). Phylogenetic diversity of microbial communities associated with the crude-oil, large-insoluble-particle and formation-water components of the reservoir fluid from a non-flooded high-temperature petroleum reservoir. Journal of bioscience and bioengineering, 113(2), 204-210.

Koonin, E. V., & Makarova, K. S. (2013). CRISPR-Cas: evolution of an RNA-based adaptive immunity system in prokaryotes. RNA biology, 10(5), 679-686.

Kumar, S., Stecher, G., & Tamura, K. (2016). MEGA7: molecular evolutionary genetics analysis version 7.0 for bigger datasets. Molecular biology and evolution, 33(7), 1870-1874.

Labrie, S. J., Samson, J. E., & Moineau, S. (2010). Bacteriophage resistance mechanisms. Nature Reviews Microbiology, 8(5), 317.

Lagesen, K., Hallin, P., Rødland, E. A., Stærfeldt, H. H., Rognes, T., & Ussery, D. W. (2007). RNAmmer: consistent and rapid annotation of ribosomal RNA genes. Nucleic acids research, 35(9), 3100-3108.

LaMontagne, M. G., & Holden, P. A. (2003). Comparison of free-living and particle-associated bacterial communities in a coastal lagoon. Microbial ecology, 46(2), 228-237.

Lander, E. S. (2016). The heroes of CRISPR. Cell, 164(1-2), 18-28.

Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. Nature methods, 9(4), 357.

Laslett, D., & Canback, B. (2004). ARAGORN, a program to detect tRNA genes and tmRNA genes in nucleotide sequences. Nucleic acids research, 32(1), 11-16.

Lau, M. C., Aitchison, J. C., & Pointing, S. B. (2009). Bacterial community composition in thermophilic microbial mats from five hot springs in central Tibet. Extremophiles, 13(1), 139-149.

Lavik, G., Stührmann, T., Brüchert, V., Van der Plas, A., Mohrholz, V., Lam, P., ... & Kuypers, M. M. (2009). Detoxification of sulphidic African shelf waters by blooming chemolithotrophs. Nature, 457(7229), 581.

Lechner, M., Findeiß, S., Steiner, L., Marz, M., Stadler, P. F., & Prohaska, S. J. (2011). Proteinortho: detection of (co-) orthologs in large-scale analysis. BMC bioinformatics, 12(1), 124.

Letunic, I., & Bork, P. (2016). Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. Nucleic acids research, 44(W1), W242-W245.

Levén, L., Eriksson, A. R., & Schnürer, A. (2007). Effect of process temperature on bacterial and archaeal communities in two methanogenic bioreactors treating organic household waste. FEMS microbiology ecology, 59(3), 683-693.

Li, D., Liu, C. M., Luo, R., Sadakane, K., & Lam, T. W. (2015). MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. Bioinformatics, 31(10), 1674-1676.

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics, 25(14), 1754-1760.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., ... & Durbin, R. (2009). The sequence alignment/map format and SAMtools. Bioinformatics, 25(16), 2078-2079.

Li, W., & Godzik, A. (2006). Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics, 22(13), 1658-1659.

Lin, Y., Cradick, T. J., Brown, M. T., Deshmukh, H., Ranjan, P., Sarode, N., ... & Bao, G. (2014). CRISPR/Cas9 systems have off-target activity with insertions or deletions between target DNA and guide RNA sequences. Nucleic acids research, 42(11), 7473-7485.

Liu, J., Wu, W., Chen, C., Sun, F., & Chen, Y. (2011). Prokaryotic diversity, composition structure, and phylogenetic analysis of microbial communities in leachate sediment ecosystems. Applied microbiology and biotechnology, 91(6), 1659-1675.

Lloyd, K. G., Schreiber, L., Petersen, D. G., Kjeldsen, K. U., Lever, M. A., Steen, A. D., ... & Schramm, A. (2013). Predominant archaea in marine sediments degrade detrital proteins. Nature, 496(7444), 215.

Lozupone, C., & Knight, R. (2005). UniFrac: a new phylogenetic method for comparing microbial communities. Applied and environmental microbiology, 71(12), 8228-8235.

Magadán, A. H., Dupuis, M. È., Villion, M., & Moineau, S. (2012). Cleavage of phage DNA by the Streptococcus thermophilus CRISPR3-Cas system. PloS one, 7(7), e40913.

Magoč, T., & Salzberg, S. L. (2011). FLASH: fast length adjustment of short reads to improve genome assemblies. Bioinformatics, 27(21), 2957-2963.

Makarova, K. S., Grishin, N. V., Shabalina, S. A., Wolf, Y. I., & Koonin, E. V. (2006). A putative RNA-interference-based immune system in prokaryotes: computational analysis of the predicted enzymatic machinery, functional analogies with eukaryotic RNAi, and hypothetical mechanisms of action. Biology direct, 1(1), 7.

Makarova, K. S., Haft, D. H., Barrangou, R., Brouns, S. J., Charpentier, E., Horvath, P., ... & Van Der Oost, J. (2011). Evolution and classification of the CRISPR–Cas systems. Nature Reviews Microbiology, 9(6), 467.

Makarova, K. S., Wolf, Y. I., & Koonin, E. V. (2013). The basic building blocks and evolution of CRISPR–Cas systems.

Mali, P., Esvelt, K. M., & Church, G. M. (2013a). Cas9 as a versatile tool for engineering biology. Nature methods, 10(10), 957.

Mali, P., Yang, L., Esvelt, K. M., Aach, J., Guell, M., DiCarlo, J. E., ... & Church, G. M. (2013b). RNA-guided human genome engineering via Cas9. Science, 339(6121), 823-826.

Marcy, Y., Ouverney, C., Bik, E. M., Lösekann, T., Ivanova, N., Martin, H. G., ... & Quake, S. R. (2007). Dissecting biological "dark matter" with single-cell genetic analysis of rare and uncultivated TM7 microbes from the human mouth. Proceedings of the National Academy of Sciences, 104(29), 11889-11894.

Marraffini, L. A. (2013). CRISPR-Cas immunity against phages: its effects on the evolution and survival of bacterial pathogens. PLoS pathogens, 9(12), e1003765.

Marraffini, L. A. (2015). CRISPR-Cas immunity in prokaryotes. Nature, 526(7571), 55.

Marraffini, L. A., & Sontheimer, E. J. (2010). CRISPR interference: RNA-directed adaptive immunity in bacteria and archaea. Nature Reviews Genetics, 11(3), 181.

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet. journal, 17(1), pp-10.

Mistry, J., Finn, R. D., Eddy, S. R., Bateman, A., & Punta, M. (2013). Challenges in homology search: HMMER3 and convergent evolution of coiled-coil regions. Nucleic acids research, 41(12), e121-e121.

Mojica, F. J. M., Diez-Villasenor, C., Garcia-Martinez, J., & Almendros, C. (2009). Short motif sequences determine the targets of the prokaryotic CRISPR defence system. Microbiology, 155(3), 733-740.

Mojica, F. J., García-Martínez, J., & Soria, E. (2005). Intervening sequences of regularly spaced prokaryotic repeats derive from foreign genetic elements. Journal of molecular evolution, 60(2), 174-182.

Moore, J. K., & Haber, J. E. (1996). Cell cycle and genetic requirements of two pathways of nonhomologous end-joining repair of double-strand breaks in Saccharomyces cerevisiae. Molecular and cellular biology, 16(5), 2164-2173.

Mussolino, C., Morbitzer, R., Lütge, F., Dannemann, N., Lahaye, T., & Cathomen, T. (2011). A novel TALE nuclease scaffold enables high genome editing activity in combination with low toxicity. Nucleic acids research, 39(21), 9283-9293.

Nam, K. H., Ding, F., Haitjema, C., Huang, Q., DeLisa, M. P., & Ke, A. (2012). Double-stranded endonuclease activity in Bacillus halodurans clustered regularly interspaced short palindromic repeats (CRISPR)-associated Cas2 protein. Journal of Biological Chemistry, 287(43), 35943-35952.

Nawrocki, E. P., Burge, S. W., Bateman, A., Daub, J., Eberhardt, R. Y., Eddy, S. R., ... & Finn, R. D. (2014). Rfam 12.0: updates to the RNA families database. Nucleic acids research, 43(D1), D130-D137.

Nawrocki, E. P., & Eddy, S. R. (2013). Infernal 1.1: 100-fold faster RNA homology searches. Bioinformatics, 29(22), 2933-2935.

Newberry, C. J., Webster, G., Cragg, B. A., Parkes, R. J., Weightman, A. J., & Fry, J. C. (2004). Diversity of prokaryotes and methanogenesis in deep subsurface sediments from the Nankai Trough, Ocean Drilling Program Leg 190. Environmental Microbiology, 6(3), 274-287.

Niemann, H., Lösekann, T., De Beer, D., Elvert, M., Nadalig, T., Knittel, K., ... & Foucher, J. P. (2006). Novel microbial communities of the Haakon Mosby mud volcano and their role as a methane sink. Nature, 443(7113), 854.

Niu, Y., Shen, B., Cui, Y., Chen, Y., Wang, J., Wang, L., ... & Xiang, A. P. (2014). Generation of gene-modified cynomolgus monkey via Cas9/RNA-mediated gene targeting in one-cell embryos. Cell, 156(4), 836-843.

Nobu, M. K., Dodsworth, J. A., Murugapiran, S. K., Rinke, C., Gies, E. A., Webster, G., ... & Jørgensen, B. B. (2016). Phylogeny and physiology of candidate phylum 'Atribacteria'(OP9/JS1) inferred from cultivation-independent genomics. The ISME journal, 10(2), 273.

Nuñez, J. K., Kranzusch, P. J., Noeske, J., Wright, A. V., Davies, C. W., & Doudna, J. A. (2014). Cas1–Cas2 complex formation mediates spacer acquisition during CRISPR–Cas adaptive immunity. Nature Structural and Molecular Biology, 21(6), 528.

Orcutt, B. N., Sylvan, J. B., Knab, N. J., & Edwards, K. J. (2011). Microbial ecology of the dark ocean above, at, and below the seafloor. Microbiology and Molecular Biology Reviews, 75(2), 361-422.

Padilla, C. C., Bertagnolli, A. D., Bristow, L. A., Sarode, N., Glass, J. B., Thamdrup, B., & Stewart, F. J. (2017). Metagenomic binning recovers a transcriptionally active Gammaproteobacterium linking methanotrophy to partial denitrification in an anoxic oxygen minimum zone. Frontiers in Marine Science, 4, 23.

Paez-Espino, D., Morovic, W., Sun, C. L., Thomas, B. C., Ueda, K. I., Stahl, B., ... & Banfield, J. F. (2013). Strong bias in the bacterial CRISPR elements that confer immunity to phage. Nature communications, 4, 1430.

Pardo, B., Gomez-Gonzalez, B., & Aguilera, A. (2009). DNA repair in mammalian cells. Cellular and Molecular Life Sciences, 66(6), 1039-1056.

Parkes, R. J., Cragg, B., Roussel, E., Webster, G., Weightman, A., & Sass, H. (2014). A review of prokaryotic populations and processes in sub-seafloor sediments, including biosphere: geosphere interactions. Marine Geology, 352, 409-425.

Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P., & Tyson, G. W. (2015). CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. Genome research, 25(7), 1043-1055.

Parsons, R. J., Breitbart, M., Lomas, M. W., & Carlson, C. A. (2012). Ocean time-series reveals recurring seasonal patterns of virioplankton dynamics in the northwestern Sargasso Sea. The ISME journal, 6(2), 273.

Pattanayak, V., Lin, S., Guilinger, J. P., Ma, E., Doudna, J. A., & Liu, D. R. (2013). High-throughput profiling of off-target DNA cleavage reveals RNA-programmed Cas9 nuclease specificity. Nature biotechnology, 31(9), 839.

Pattanayak, V., Ramirez, C. L., Joung, J. K., & Liu, D. R. (2011). Revealing off-target cleavage specificities of zinc-finger nucleases by in vitro selection. Nature methods, 8(9), 765.

Paulmier, A., & Ruiz-Pino, D. (2009). Oxygen minimum zones (OMZs) in the modern ocean. Progress in Oceanography, 80(3-4), 113-128.

Peacock, J. P., Cole, J. K., Murugapiran, S. K., Dodsworth, J. A., Fisher, J. C., Moser, D. P., & Hedlund, B. P. (2013). Pyrosequencing reveals high-temperature cellulolytic microbial consortia in Great Boiling Spring after in situ lignocellulose enrichment. PLoS One, 8(3), e59927.

Pham, V. D., Hnatow, L. L., Zhang, S., Fallon, R. D., Jackson, S. C., Tomb, J. F., ... & Keeler, S. J. (2009). Characterizing microbial diversity in production water from an Alaskan mesothermic petroleum reservoir with two independent molecular methods. Environmental microbiology, 11(1), 176-187.

Ploug, H., Grossart, H. P., Azam, F., & Jørgensen, B. B. (1999). Photosynthesis, respiration, and carbon turnover in sinking marine snow from surface waters of Southern California Bight: implications for the carbon cycle in the ocean. Marine Ecology Progress Series, 1-11.

Pourcel, C., Salvignol, G., & Vergnaud, G. (2005). CRISPR elements in Yersinia pestis acquire new repeats by preferential uptake of bacteriophage DNA, and provide additional tools for evolutionary studies. Microbiology, 151(3), 653-663.

Proctor, L. M., & Fuhrman, J. A. (1990). Viral mortality of marine bacteria and cyanobacteria. Nature, 343(6253), 60.

Ptashne, M. (2004). A genetic switch: phage lambda revisited. CSHL press.

Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., ... & Glöckner, F. O. (2012). The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic acids research, 41(D1), D590-D596.

Ramirez, C. L., Certo, M. T., Mussolino, C., Goodwin, M. J., Cradick, T. J., McCaffrey, A. P., ... & Joung, J. K. (2012). Engineered zinc finger nickases induce homology-directed repair with reduced mutagenic effects. Nucleic acids research, 40(12), 5560-5568.

Ran, F. A., Cong, L., Yan, W. X., Scott, D. A., Gootenberg, J. S., Kriz, A. J., ... & Koonin, E. V. (2015). In vivo genome editing using Staphylococcus aureus Cas9. Nature, 520(7546), 186.

Reed, D. W., Fujita, Y., Delwiche, M. E., Blackwelder, D. B., Sheridan, P. P., Uchida, T., & Colwell, F. S. (2002). Microbial communities from methane hydrate-bearing deep marine sediments in a forearc basin. Applied and Environmental Microbiology, 68(8), 3759-3770.

Reeks, J., Naismith, J. H., & White, M. F. (2013). CRISPR interference: a structural perspective. Biochemical Journal, 453(2), 155-166.

Rinke, C., Schwientek, P., Sczyrba, A., Ivanova, N. N., Anderson, I. J., Cheng, J. F., ... & Dodsworth, J. A. (2013). Insights into the phylogeny and coding potential of microbial dark matter. Nature, 499(7459), 431.

Riviere, D., Desvignes, V., Pelletier, E., Chaussonnerie, S., Guermazi, S., Weissenbach, J., ... & Sghir, A. (2009). Towards the definition of a core of microorganisms involved in anaerobic digestion of sludge. The ISME journal, 3(6), 700.

Rodriguez-R, L. M., & Konstantinidis, K. T. (2016). The enveomics collection: a toolbox for specialized analyses of microbial genomes and metagenomes. PeerJ Preprints.

Samai, P., Pyenson, N., Jiang, W., Goldberg, G. W., Hatoum-Aslan, A., & Marraffini, L. A. (2015). Co-transcriptional DNA and RNA cleavage during type III CRISPR-Cas immunity. Cell, 161(5), 1164-1174.

Samai, P., Smith, P., & Shuman, S. (2010). Structure of a CRISPR-associated protein Cas2 from Desulfovibrio vulgaris. Acta Crystallographica Section F: Structural Biology and Crystallization Communications, 66(12), 1552-1556.

Sander, J. D., & Joung, J. K. (2014). CRISPR-Cas systems for editing, regulating and targeting genomes. Nature biotechnology, 32(4), 347.

Sapranauskas, R., Gasiunas, G., Fremaux, C., Barrangou, R., Horvath, P., & Siksnys, V. (2011). The Streptococcus thermophilus CRISPR/Cas system provides immunity in Escherichia coli. Nucleic acids research, 39(21), 9275-9282.

Sashital, D. G., Wiedenheft, B., & Doudna, J. A. (2012). Mechanism of foreign DNA selection in a bacterial adaptive immune system. Molecular cell, 46(5), 606-615.

Sayeh, R., Birrien, J. L., Alain, K., Barbier, G., Hamdi, M., & Prieur, D. (2010). Microbial diversity in Tunisian geothermal springs as detected by molecular and culture-based approaches. Extremophiles, 14(6), 501-514.

Schmieder, R., & Edwards, R. (2011). Quality control and preprocessing of metagenomic datasets. Bioinformatics, 27(6), 863-864.

Scranton, M. I., Astor, Y., Bohrer, R., Ho, T. Y., & Muller-Karger, F. (2001). Controls on temporal variability of the geochemistry of the deep Cariaco Basin. Deep Sea Research Part I: Oceanographic Research Papers, 48(7), 1605-1625.

Seemann, T. (2014). Prokka: rapid prokaryotic genome annotation. Bioinformatics, 30(14), 2068-2069.

Selle, K., & Barrangou, R. (2015). Harnessing CRISPR–Cas systems for bacterial genome editing. Trends in microbiology, 23(4), 225-232.

Semenova, E., Jore, M. M., Datsenko, K. A., Semenova, A., Westra, E. R., Wanner, B., ... & Severinov, K. (2011). Interference by clustered regularly interspaced short palindromic repeat (CRISPR) RNA is governed by a seed sequence. Proceedings of the National Academy of Sciences, 108(25), 10098-10103.

Sogin, M. L., Morrison, H. G., Huber, J. A., Welch, D. M., Huse, S. M., Neal, P. R., ... & Herndl, G. J. (2006). Microbial diversity in the deep sea and the underexplored

"rare biosphere". Proceedings of the National Academy of Sciences, 103(32), 12115-12120.

Sorek, R., Lawrence, C. M., & Wiedenheft, B. (2013). CRISPR-mediated adaptive immune systems in bacteria and archaea. Annual review of biochemistry, 82, 237-266.

Sorokin, V. A., Gelfand, M. S., & Artamonova, I. I. (2010). Evolutionary dynamics of clustered irregularly interspaced short palindromic repeat systems in the ocean metagenome. Applied and environmental microbiology, 76(7), 2136-2144.

Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics, 30(9), 1312-1313.

Stevens, H., & Ulloa, O. (2008). Bacterial diversity in the oxygen minimum zone of the eastern tropical South Pacific. Environmental microbiology, 10(5), 1244-1259.

Stewart, F. J., Ulloa, O., & DeLong, E. F. (2012). Microbial metatranscriptomics in a permanent marine oxygen minimum zone. Environmental microbiology, 14(1), 23-40.

Stocker, R. (2012). Marine microbes see a sea of gradients. science, 338(6107), 628-633.

Stramma, L., Johnson, G. C., Sprintall, J., & Mohrholz, V. (2008). Expanding oxygen-minimum zones in the tropical oceans. science, 320(5876), 655-658.

Sugimoto, N., Nakano, S. I., Katoh, M., Matsumura, A., Nakamuta, H., Ohmichi, T., ... & Sasaki, M. (1995). Thermodynamic parameters to predict stability of RNA/DNA hybrid duplexes. Biochemistry, 34(35), 11211-11216.

Sun, C. L., Barrangou, R., Thomas, B. C., Horvath, P., Fremaux, C., & Banfield, J. F. (2013). Phage mutations in response to CRISPR diversification in a bacterial population. Environmental microbiology, 15(2), 463-470.

Suttle, C. A. (2005). Viruses in the sea. Nature, 437(7057), 356.

Suttle, C. A. (2007). Marine viruses—major players in the global ecosystem. Nature Reviews Microbiology, 5(10), 801.

Tang, Y. Q., Ji, P., Hayashi, J., Koike, Y., Wu, X. L., & Kida, K. (2011). Characteristic microbial community of a dry thermophilic methanogenic digester: its long-term stability and change with feeding. Applied microbiology and biotechnology, 91(5), 1447.

Terns, R. M., & Terns, M. P. (2014). CRISPR-based technologies: prokaryotic defense weapons repurposed. Trends in Genetics, 30(3), 111-118.

Teske, A., Hinrichs, K. U., Edgcomb, V., de Vera Gomez, A., Kysela, D., Sylva, S. P., ... & Jannasch, H. W. (2002). Microbial diversity of hydrothermal sediments in the Guaymas Basin: evidence for anaerobic methanotrophic communities. Applied and Environmental Microbiology, 68(4), 1994-2007.

Tesson, L., Usal, C., Ménoret, S., Leung, E., Niles, B. J., Remy, S., ... & Gregory, P. D. (2011). Knockout rats generated by embryo microinjection of TALENs. Nature biotechnology, 29(8), 695.

Thamdrup, B., Dalsgaard, T., & Revsbech, N. P. (2012). Widespread functional anoxia in the oxygen minimum zone of the Eastern South Pacific. Deep Sea Research Part I: Oceanographic Research Papers, 65, 36-45.

Ulloa, O., Canfield, D. E., DeLong, E. F., Letelier, R. M., & Stewart, F. J. (2012). Microbial oceanography of anoxic oxygen minimum zones. Proceedings of the National Academy of Sciences, 109(40), 15996-16003.

Ulloa, O., & Pantoja, S. (2009). The oxygen minimum zone of the eastern South Pacific. Deep Sea Research Part II: Topical Studies in Oceanography, 56(16), 987-991.

Van der Oost, J., Jore, M. M., Westra, E. R., Lundgren, M., & Brouns, S. J. (2009). CRISPR-based adaptive and heritable immunity in prokaryotes. Trends in biochemical sciences, 34(8), 401-407.

van Duijn, E., Barbu, I. M., Barendregt, A., Jore, M. M., Wiedenheft, B., Lundgren, M., ... & Heck, A. J. (2012). Native tandem and ion mobility mass spectrometry highlight structural and modular similarities in clustered-regularly-interspaced shot-palindromic-repeats (CRISPR)-associated protein complexes from Escherichia coli and Pseudomonas aeruginosa. Molecular & Cellular Proteomics, 11(11), 1430-1441.

Vick, T. J., Dodsworth, J. A., Costa, K. C., Shock, E. L., & Hedlund, B. P. (2010). Microbiology and geochemistry of Little Hot Creek, a hot spring environment in the Long Valley Caldera. Geobiology, 8(2), 140-154.

Vik, D. R., Roux, S., Brum, J. R., Bolduc, B., Emerson, J. B., Padilla, C. C., ... & Sullivan, M. B. (2017). Putative archaeal viruses from the mesopelagic ocean. PeerJ, 5, e3428.

Walsh, D. A., Zaikova, E., Howes, C. G., Song, Y. C., Wright, J. J., Tringe, S. G., ... & Hallam, S. J. (2009). Metagenome of a versatile chemolithoautotroph from expanding oceanic dead zones. Science, 326(5952), 578-582.

Wang, R., Preamplume, G., Terns, M. P., Terns, R. M., & Li, H. (2011). Interaction of the Cas6 riboendonuclease with CRISPR RNAs: recognition and cleavage. Structure, 19(2), 257-264.

Webster, G., Parkes, R. J., Cragg, B. A., Newberry, C. J., Weightman, A. J., & Fry, J. C. (2006). Prokaryotic community composition and biogeochemical processes in deep subseafloor sediments from the Peru Margin. FEMS microbiology ecology, 58(1), 65-85.

Webster, G., Parkes, R. J., Fry, J. C., & Weightman, A. J. (2004). Widespread occurrence of a novel division of bacteria identified by 16S rRNA gene sequences originally found in deep marine sediments. Applied and Environmental Microbiology, 70(9), 5708-5713.

Webster, G., Yarram, L., Freese, E., Köster, J., Sass, H., Parkes, R. J., & Weightman, A. J. (2007). Distribution of candidate division JS1 and other Bacteria in tidal sediments of the German Wadden Sea using targeted 16S rRNA gene PCR-DGGE. FEMS microbiology ecology, 62(1), 78-89.

Weinbauer, M. G., Fuks, D., & Peduzzi, P. (1993). Distribution of viruses and dissolved DNA along a coastal trophic gradient in the northern Adriatic Sea. Applied and Environmental Microbiology, 59(12), 4074-4082.

Wemheuer B., Taube R., Akyol P., Wemheuer F., Daniel R. (2013) Microbial Diversity and Biochemical Potential Encoded by Thermal Spring Metagenomes Derived from the Kamchatka Peninsula. Archaea, 2013, 13.

Westra, E. R., Swarts, D. C., Staals, R. H., Jore, M. M., Brouns, S. J., & van der Oost, J. (2012). The CRISPRs, they are a-changin': how prokaryotes generate adaptive immunity. Annual review of genetics, 46, 311-339.

Wiedenheft, B., Lander, G. C., Zhou, K., Jore, M. M., Brouns, S. J., van der Oost, J., ... & Nogales, E. (2011). Structures of the RNA-guided surveillance complex from a bacterial immune system. Nature, 477(7365), 486.

Wiedenheft, B., Sternberg, S. H., & Doudna, J. A. (2012). RNA-guided genetic silencing systems in bacteria and archaea. Nature, 482(7385), 331.

Wilms, R., Köpke, B., Sass, H., Chang, T. S., Cypionka, H., & Engelen, B. (2006). Deep biosphere-related bacteria within the subsurface of tidal flat sediments. Environmental Microbiology, 8(4), 709-719.

Wommack, K. E., & Colwell, R. R. (2000). Virioplankton: viruses in aquatic ecosystems. Microbiology and molecular biology reviews, 64(1), 69-114.

Wright, J. J., Konwar, K. M., & Hallam, S. J. (2012). Microbial ecology of expanding oxygen minimum zones. Nature Reviews Microbiology, 10(6), 381.

Wrighton, K. C., Agbo, P., Warnecke, F., Weber, K. A., Brodie, E. L., DeSantis, T. Z., ... & Coates, J. D. (2008). A novel ecological role of the Firmicutes identified in thermophilic microbial fuel cells. The ISME journal, 2(11), 1146.

Wyrtki, K. (1962, January). The oxygen minima in relation to ocean circulation. In Deep Sea Research and Oceanographic Abstracts (Vol. 9, No. 1-2, pp. 11-23). Elsevier.

Yarza, P., Yilmaz, P., Pruesse, E., Glöckner, F. O., Ludwig, W., Schleifer, K. H., ... & Rosselló-Móra, R. (2014). Uniting the classification of cultured and uncultured bacteria and archaea using 16S rRNA gene sequences. Nature Reviews Microbiology, 12(9), 635.

Ye, L., Wang, J., Beyer, A. I., Teque, F., Cradick, T. J., Qi, Z., ... & Levy, J. A. (2014). Seamless modification of wild-type induced pluripotent stem cells to the natural CCR5Δ32 mutation confers resistance to HIV infection. Proceedings of the National Academy of Sciences, 111(26), 9591-9596.

Yosef, I., Goren, M. G., & Qimron, U. (2012). Proteins and DNA elements essential for the CRISPR adaptation process in Escherichia coli. Nucleic acids research, 40(12), 5569-5576.

Zhang, J., Rouillon, C., Kerou, M., Reeks, J., Brugger, K., Graham, S., ... & Naismith, J. H. (2012). Structure and mechanism of the CMR complex for CRISPR-mediated antiviral immunity. Molecular cell, 45(3), 303-313.

Zhen, S., Hua, L., Liu, Y. H., Gao, L. C., Fu, J., Wan, D. Y., ... & Gao, X. (2015). Harnessing the clustered regularly interspaced short palindromic repeat (CRISPR)/CRISPR-associated Cas9 system to disrupt the hepatitis B virus. Gene therapy, 22(5), 404.